

Glycoprotein and Glycopeptide Analysis by Liquid Chromatography and Mass Spectrometry

By

Zhikai Zhu

Submitted to the graduate degree program in the Department of Chemistry
and the Graduate Faculty of the University of Kansas in partial fulfillment of
the requirements for the degree of Doctor of Philosophy.

Chairperson: Dr. Heather Desaire,

Dr. Cindy L. Berrie

Dr. Minae Mure

Dr. Yong Zeng

Dr. Josephine R. Chandler

Date Defended: July 7th, 2015

The Dissertation Committee for **Zhikai Zhu**
certifies that this is the approved version of the following dissertation:

**Glycoprotein and Glycopeptide Analysis by Liquid Chromatography and
Mass Spectrometry**

Chairperson: Heather Desaire, Ph. D.

Date approved: July 7th, 2015

ABSTRACT

Carbohydrates on proteins play essential roles in mediating various biological activities, such as cell adhesion, cell signaling, and antigen-antibody recognition. Altered glycosylation profiles on proteins have been found to be closely related to the progression of certain diseases including cancer. Moreover, the majority of biological therapeutics are glycoproteins, and variations in their glycosylation states would impact the efficacy of these pharmaceuticals. In order to gain in-depth understanding of the structure and function relationship of glycoproteins, the glycosylation profile associated with the protein needs to be determined. The workflow for site-specific glycosylation analysis involves the enzymatic digestion of glycoproteins, and the resulting glycopeptides are analyzed by LC-MS and tandem MS. Among different fragmentation modes in tandem MS, electron transfer dissociation (ETD) is extremely useful in revealing the sequence and glycan location of a glycopeptide. However, data analysis is a huge bottleneck for high throughput glycopeptide identifications using ETD. In this dissertation, this analytical challenge is addressed in multiple facets. Firstly, ETD-MS/MS data of N-linked glycopeptides with known peptide sequences and glycan compositions are collected to build a training dataset. By studying the training dataset, the fragmentation patterns of glycopeptides are summarized to develop an effective algorithm for scoring of the glycopeptide ETD data. A software tool is built based on the algorithm to interpret data collected from a clade C HIV envelope glycoprotein, gp140, and no false positive assignment is made by the program. Secondly, the fragmentation of O-linked glycopeptides in ETD is systematically studied, such that useful rules are found to facilitate O-glycopeptide identifications. The rules are implemented into an algorithm to score O-glycopeptide candidates against the ETD data, and the developed algorithm demonstrates superior performance compared to a publicly available data analysis tool. Lastly, a new

glycoproteomics software is outlined to evaluate the false discovery rate of automated glycopeptide assignments reported by a computer program. In sum, this dissertation advances the glycoproteomics field by establishing an integral system for expedited glycopeptide analysis with improved accuracy. The other part of the dissertation details an absolute quantitation approach for determining the extent of glycosylation on individual glycosylation sites on a protein, and the developed method quantifies the glycosylation site occupancy more accurately than the conventional approach.

ACKNOWLEDGEMENTS

I owe a lot to many great people in my life, and during the course of my graduate study, my advisor, Dr. Heather Desaire, is certainly the first and most important figure that I am really appreciative of. Her encouragement and guidance persist even before my arrival in the United States, and her mentoring is, and will always be the torch that enlightens me when I meet difficulties, being it in research or in life. I feel extremely lucky to join the Desaire group, and have worked with some excellent scientists from here. Dr. Eden Go, thank you so much for taking the time to train me as a rookie, bearing with me the clumsiness, and helping me to make improvement towards being a real mass spectrometrists. I have always enjoyed the fun conversations with you about science and everything else. The soon-to-be Dr. Dan Clark, thank you so much for sharing your in-depth knowledge in chemistry and instrumentation with me, and the archery shooting game was very interesting. I would also like to thank the past and present Desaire group members for their sincere help and support, including David, Dr. Katie Rebecchi, Dr. Carrie Woodin, Xiaomeng, Jude, Kasun and Wenting, to name a few.

I would like to express my deep gratitude to Dr. Minae Mure, Dr. Cindy Berrie, Dr. Yong Zeng, and Dr. Josephine Chandler for being my dissertation committee members. Their insights and suggestions greatly improve the quality of the dissertation.

I would also like to thank some of the friends that I have met in Lawrence, including Yan, Qiqiang, Len Andyshack and Yunan, for their kindly help; Dr. Hongbin Liu from Genentech has also served as a great mentor to me to unveil some of the best research in biotechnology industry.

Thank you to my mother and father, Mrs. Huiping Zhou and Mr. Junmin Zhu, for the great love and support throughout my life. The gratitude is beyond words and I am very blessed to have your company. I would also like to thank my entire family in China for their bonding love.

Thank you to all the friends and people that have helped me in various ways during my Ph.D. study.

Finally, thank you my Lord, for being with me and showing me the right path in life. I will follow you for the rest of my life and beyond that.

TABLE OF CONTENTS

CHAPTER I	1
INTRODUCTION	
1.1 Glycosylation on Proteins	1
1.1.1 N-Linked Glycosylation	1
1.1.2 Glycosylation Site Occupancy	2
1.1.3 O-Linked Glycosylation	3
1.1.4 Site-Specific Glycosylation Analysis	4
1.2 Glycoprotein Purification and Digestion	6
1.2.1 Glycoprotein/Glycopeptide Enrichment	6
1.2.2 Glycoprotein Digestion	8
1.3 Analytical Separation of Glycopeptides	10
1.4 Mass Spectrometric Analysis of Glycopeptides	12
1.4.1 ESI and MALDI Mass Analysis	12
1.4.2 Tandem MS of Glycopeptides	13
1.4.2.1 Collisional Dissociation Methods	13
1.4.2.2 Electron Based Fragmentation	15
1.4.2.3 Photon-Induced Dissociation	17
1.4.3 Ion Mobility Spectrometry	19
1.5 Automated Glycoproteomics	21
1.6 Conclusions	24
1.7 Summary of Subsequent Chapters	26
References	29

GlycoPep Detector: A tool for assigning N-linked glycopeptides based on their ETD-MS/MS spectra

2.1 Introduction	38
2.2 Experimental Procedure	40
2.2.1 Materials and Reagents	40
2.2.2 Glycoprotein Digestion	40
2.2.3 Glycopeptide Enrichment	40
2.2.4 Direct Injection Mass Spectrometry	41
2.2.5 Liquid Chromatography/Mass Spectrometry	41
2.2.6 Glycopeptide Training Dataset	42
2.2.7 Glycopeptide Validation Dataset	43
2.2.8 Glycopeptide Data Input	43
2.2.9 False Positive Rate Determination	44
2.3 Results and Discussion	45
2.3.1 Spectral Library Generation	45
2.3.2 Characteristics of Glycopeptide Fragmentation in ETD	45
2.3.3 Data Pre-processing of the Raw Spectrum	50
2.3.4 Novel Aspects of GPD Algorithm	52
2.3.5 GPD Scoring of ETD Spectra from Model Glycoproteins	53
2.3.6 Extended Test of GPD on ETD Spectra from the HIV Env Glycoprotein, C.97ZA012 gp140ΔCFI	57
2.4 Conclusion	59

References	60
------------	----

CHAPTER III	92
--------------------	-----------

**Characterizing O-linked glycopeptides by electron transfer dissociation:
fragmentation rules and applications in data analysis**

3.1 Introduction	93
3.2 Experimental Procedures	95
3.2.1 Samples and Reagents	95
3.2.2 Sample Preparation	95
3.2.3 LC-MS/MS	96
3.2.4 Direct-infusion MS/MS	96
3.2.5 Data Analysis	97
3.2.6 Algorithm Performance Test	97
3.3 Results and Discussion	98
3.3.1 O-Glycopeptide Fragmentation Rules in ETD	99
3.3.2 Algorithm Design and Implementation in O-glycopeptide ETD Data Analysis	105
3.3.3 Algorithm Scoring of O-glycopeptide Candidate Compositions	106
3.3.4 Analysis of O-linked Glycopeptide ETD Data Sets by GlycoPep Scorer	109
3.4 Conclusion	112

References	113
------------	-----

Determination of the False Discovery Rate in N-Linked Glycopeptide Identifications by GlycoPep Evaluator

4.1 Introduction	132
4.2 Experimental Procedures	135
4.2.1 Samples and Reagents	135
4.2.2 Protease Digestion	135
4.2.3 LC/MS Analysis	135
4.2.4 Glycopeptide MS/MS Data Set	136
4.2.5 Decoy and Target Candidates Generation	137
4.2.6 Scoring of Decoy and Target Candidates	138
4.2.7 False Discovery Rate Study Using GPE	138
4.3 Results and Discussion	138
4.3.1 Overview of GlycoPep Evaluator	139
4.3.2 False Discovery Rate Analysis	142
4.3.3 Target and Decoy Glycopeptides Analysis	143
4.3.4 Is GPE Consistently Able To Identify The Correct Candidate, When It Is Present?	147
4.3.5 Is GPE Effective At Identifying Misassigned Spectra?	148
4.3.6 Comparison of the predicted FDR to the true FDR	148
4.4 Conclusions	151
References	153

Absolute Quantitation of Glycosylation Site Occupancy Using Isotopically Labeled Standards and LC-MS

5.1 Introduction	158
5.2 Experimental Procedures	160
5.2.1 Materials and Reagents	160
5.2.2 Sample Preparation	160
5.2.3 N-deglycosylation	161
5.2.4 LC-MS Analysis	161
5.3 Results and Discussion	162
5.4 Conclusions	168
References	170

LIST OF TABLES

CHAPTER 2

Table 1	GPD scoring parameters of the correct glycopeptide composition and a decoy candidate	54
Table S1	GPD scoring list of model glycopeptides based on the training dataset.	63
Table S2	GPD scoring list of glycopeptides from HIV Env Glycoprotein based on the validation dataset	73

CHAPTER 3

Table 1	Algorithm scoring results of the ETD-MS/MS data against 10 <i>O</i> -glycopeptide compositions.	108
Table 2	Analysis summary of GlycoPep Scorer and Protein Prospector in interpreting <i>O</i> -glycopeptide ETD datasets	111
Table S1	Score results by GlycoPep Scorer for <i>O</i> -linked glycopeptide ETD data set	116
Table S2	Score results by Protein Prospector for <i>O</i> -glycopeptide ETD data set	122

CHAPTER 4

Table S1	The decoy glycopeptides and their scoring results returned by GPE for the four incorrect glycopeptides that were input as target candidates	156
-----------------	---	-----

CHAPTER 5

Table S1	Algorithm scoring results of the ETD-MS/MS data against 10 <i>O</i> -glycopeptide compositions.	172
Table S2	Analysis summary of GlycoPep Scorer and Protein Prospector in interpreting <i>O</i> -glycopeptide ETD datasets	173

LIST OF FIGURES

CHAPTER 1

Figure 1-1	Symbolic representations of N-linked glycan structures	3
Figure 1-2	The eight core structures of mucin-type O-linked glycans	4
Figure 1-3	A general diagram of the glycopeptide-based analysis workflow by mass spectrometry	5
Figure 1-4	Extracted ion chromatograms of glycopeptides separated by a PGC LC-chip	12
Figure 1-5	CID-MS/MS data of a biantennary complex N-linked glycopeptide from trypsin digested transferrin	15
Figure 1-6	ETD data of a glycopeptide carrying a triantennary complex N-glycan at different charge states: (A) 4+, (B) 5+, (C) 6+.	17
Figure 1-7	UVPD data of a doubly deprotonated O-glycopeptide anion from kappa casein, whose structure is depicted in the figure	19
Figure 1-8	Glycopeptide analysis by quadrupole-ion mobility-TOF MS	21

CHAPTER 2

Figure 2-1	ETD-MS/MS data of a 32-amino acid-long glycopeptide with a tri-antennary complex <i>N</i> -glycan of asialofetuin at 4+ (A, m/z	46
-------------------	---	----

1415.6), 5+ (B, m/z 1132.7) and 6+ (C, m/z 944.1) charge states

Figure 2-2	The eight core structures of mucin-type O-linked glycans	48
Figure 2-3	ETD data from two avidin glycopeptides at 3+ charge state with the same peptide sequence and homologous high mannose <i>N</i> -glycans attached: Man7 (A, m/z 1126.1) and Man6 (B, m/z 1072.1) tryptic glycopeptides.	49
Figure 2-4	Scheme of the spectral pre-processing method	51
Figure 2-5	CID-MS/MS data of a biantennary complex N-linked glycopeptide from trypsin digested transferrin	56
Figure 2-6	ETD data of a glycopeptide carrying a triantennary complex <i>N</i> -glycan at different charge states: (A) 4+, (B) 5+, (C) 6+.	57
Figure 2-7	A summary of the final scores of 45 distinct glycopeptides from the HIV Env Glycoprotein (shown in red bars).	58

CHAPTER 3

Figure 3-1	ETD-MS/MS data from (A) a mucin <i>O</i> -linked glycopeptide of which the glycan is attached to the Thr-3 residue (2+, m/z 852.9); (B) an isomeric <i>O</i> -glycopeptide of (A) that has the same composition but with a different modification site at Thr-13 (2+, m/z 852.9); (C) a doubly charged <i>O</i> -glycopeptide from erythropoietin with the Ser-10 residue glycosylated (m/z 834.9); (D) the same glycopeptide as (C) but at 3+ charge state (m/z 557.0).	100
-------------------	--	-----

Figure 3-2	ETD spectra of (A) an <i>O</i> -linked core-2-type glycopeptide (3+, m/z 955.1) and (B) a core-1-type <i>O</i> -glycopeptide (4+, m/z 696.6) from the HIV envelope glycoprotein, and (C) a hybrid-type <i>N</i> -linked glycopeptide (3+, 1153.5) from avidin	102
-------------------	---	-----

Figure 3-3	(A) ETD-MS/MS data of an <i>O</i> -linked glycopeptide (3+, m/z 833.4) with its composition shown in (B), note that two enlarged windows showing the zoomed m/z regions of 300-500 and of 1150-1350, respectively are also present in the figure; (B) processed ETD data of (A) after spectral filtering to remove noise peaks, and the inset table lists the scoring results (including the individual ion series' scores and the respective weightings) of the correct glycopeptide composition against the processed data; (C) CID-MS/MS data of the same glycopeptide as shown in (B).	104
-------------------	--	-----

CHAPTER 4

Figure 4-1	(A) The graphical user interface (GUI) of the GlycoPep Evaluator (GPE) program. (B) The result of decoy generation completed by GPE that contains the input target glycopeptide as well as 20 decoy glycopeptide sequences generated by the program	141
-------------------	---	-----

Figure 4-2	(A) ETD-MS/MS data of a HIV gp140 glycopeptide that has a core-fucosylated biantennary complex-type glycan as shown in the figure. The peptide backbone fragment ions (c- and z-ions) are labeled. (B) CID data of the same glycopeptide in (A)	146
-------------------	---	-----

Figure 4-3	For the input glycopeptide composition (labeled as target) DGGEDNKTEEIFRPGGGNMK + [Hex]3[HexNAc]4[Fuc]1, 20 decoy glycopeptide compositions were generated by GPE. Subsequently, GPE scored both the target and decoy glycopeptides against the ETD data, and they were ranked from high to low score	147
-------------------	---	-----

as shown in this figure

Figure 4-4	Lines that are fitted based on the blue data points: correlation curves between the predicted FDR values calculated using our method, and the observed FDR values that are manually verified	150
-------------------	--	-----

CHAPTER 5

Figure 5-1	(a) Mean values of percent site occupancy (n=3) calculated by using the measured concentration of three different peptides (P1-3) and of the non-glycosylated peptide (P4) in a fetuin solution of 0.6 $\mu\text{g}/\mu\text{L}$; the rightmost bar indicates the site occupancy (n=5) determined by using PNGase to deglycosylate the protein. (b) The percent site occupancy shown was measured in the same way as (a), but herein the analyzed fetuin samples were incompletely digested.	166
-------------------	---	-----

Figure 5-2	(a) and (b) Zoomed-in mass spectra at retention times of 51.0-51.2 min and 49.5-49.7 min, respectively, of the digested fetuin solution containing heavy isotope labeled standards, with no PNGase F added. (c) and (d) CID-MS/MS data of the non-glycosylated peptide (P4) and its chemical deamidation product (deamidated P4), respectively. In (c), the potential glycosylation site Asn-158 and fragment ions that contain the site are labeled in blue; in (d), fragment ions that contain the glycosylation site are labeled in red.	168
-------------------	---	-----

CHAPTER I

Introduction

1.1 Glycosylation on Proteins

Carbohydrates can be covalently attached to the side chains of certain amino acids on proteins after their translation in the ribosome.¹⁻³ This process, referred to as protein glycosylation, is commonly observed in eukaryotic cells, and previous reports estimate that a significant portion of all proteins are glycoproteins.⁴⁻⁵ Glycosylation is mediated by enzymes found in the endoplasmic reticulum and Golgi apparatus, with glycans of diverse compositions and linkages built on the basis of individual monosaccharide units.⁶⁻⁷ These glycans (also called oligosaccharides) greatly impact the functionality of glycoproteins that participate in numerous biological events. It is widely known that glycans on membrane proteins play a major role in cellular recognition and signaling;⁸ changes in glycosylation patterns of certain endogenous proteins are intrinsically related to disease progression, as evidenced by differential protein glycosylation in congenital disorders of glycosylation and cancer.⁹⁻¹⁰ Additionally, many pharmaceuticals, such as monoclonal antibodies, are glycoprotein drugs; and variations in glycosylation during manufacturing processes can affect the stability, potency and safety of these medicines.¹¹ Therefore, protein glycosylation analysis is not only attractive as a subject of bioanalytical chemistry but is important to other areas including life sciences and biomedicine.

1.1.1 N-Linked Glycosylation

As the major glycosylation type, N-linked glycosylation got its name in that the glycans are linked through the nitrogen atom on the side-chain of an asparagine (Asn) residue.¹² It has been found that the glycosylated Asn is within a consensus sequence of Asn-Xxx-Ser/Thr (where Xxx

can be any amino acid except proline), and some evidence suggests that the amino acid at the Ser/Thr position is cysteine for a few glycoproteins.¹³

During the initial stage of the glycosylation pathway, a precursor glycan of $\text{Glc}_3\text{Man}_9\text{GlcNAc}_2$ (Figure 1) is attached to the protein, and it is then subjected to step-wise enzymatic trimming and modification so that glycans of different structures can be eventually formed. Nonetheless, all the N-linked glycans share the same conserved core of $\text{Man}_3\text{GlcNAc}_2$ (Figure 1), which is the remaining part of the precursor glycan. According to their structural features, N-glycans are divided into three groups: high-mannose, complex, and hybrid; their representative structures are shown in Figure 1. High-mannose glycans are formed by removal of the monosaccharides in the precursor glycan with no addition of any new carbohydrate, and only mannoses (Man) are left to be linked to the conserved core. By contrast, complex-type glycans are generated by both removal and addition of sugar units, with N-acetylglucosamine (GlcNAc) monosaccharides directly attached to the $\text{Man}_3\text{GlcNAc}_2$ core. The GlcNAc branch may be further extended with galactose (Gal) and terminated with N-acetylneuraminic acid (Neu5Ac, also called sialic acid). In addition, fucose (Fuc) may be added to the branching GlcNAc or the GlcNAc at the reducing-end of the conserved core. The third N-glycan group, hybrid-type glycans, contains building blocks from both high-mannose and complex-type glycans (Figure 1).

1.1.2 Glycosylation Site Occupancy

The extent of N-linked glycosylation on proteins may not always be 100%, as the biosynthesis of the carbohydrates largely depends on the enzyme that catalyzes this process. Underglycosylation leads to many diseases in plants and animals; consequently, glycosylation site occupancy is an important parameter that needs to be accurately determined.¹⁴⁻¹⁵ In order to measure the N-glycosylation site occupancy, an enzyme called protein N-glycosidase F (PNGase

F) is typically used to convert the glycosylated asparagine (N) to aspartic acid (D). By doing this conversion, the formerly glycosylated asparagine (now as aspartic acid) can be directly compared to the nonglycosylated asparagine, and the glycosylation site occupancy is thus determined. Nonetheless, this methodology has the pitfall that spontaneous chemical deamidation, during which a proportion of the nonglycosylated asparagine is converted to aspartic acid, would lead to inaccurate estimation of the site occupancy value.¹⁶⁻¹⁷

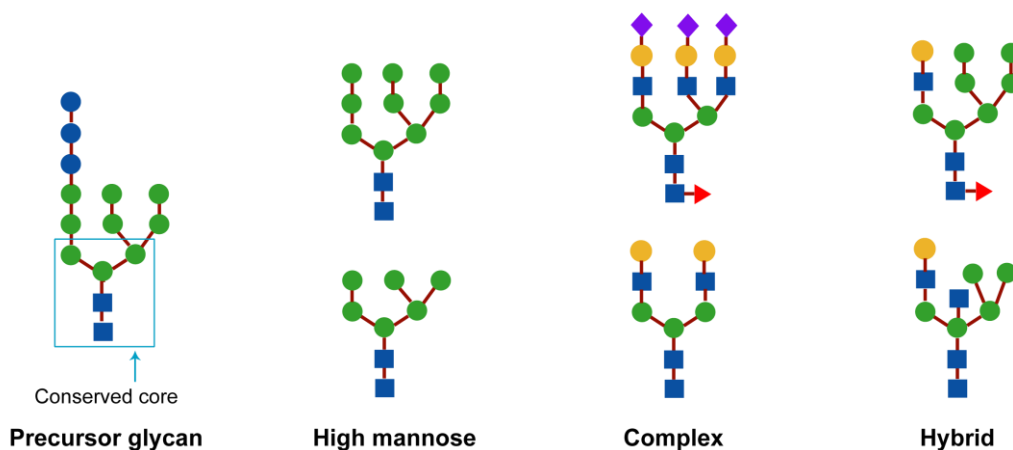


Figure 1. Symbolic representations of N-linked glycan structures. Glycan symbols include, blue square: N-acetylglucosamine (GlcNAc), blue circle: glucose (Glc), green circle: mannose (Man), yellow circle: galactose (Gal), purple diamond: N-acetylneuraminic acid (Neu5Ac), red triangle: fucose (Fuc). Following figures use the same symbols unless otherwise noted.

1.1.3 O-Linked Glycosylation

O-linked glycosylation is another widely seen glycosylation type on proteins, where glycans are bonded to the hydroxyl groups on side-chains of serine (Ser) and threonine (Thr) residues.¹⁸ Different from N-glycosylation, O-glycosylation has neither a consensus sequence that predicts the glycosylation site nor a conserved core in the glycan structure. Instead, various carbohydrate moieties such as mannose (Man), fucose (Fuc) and glucose (Glc), have been found to be bonded directly to Ser/Thr, and further attachment or modification can be made based on these core monosaccharides to generate different O-linked glycans.¹⁸⁻¹⁹ A single β -N-acetylglucosamine

(GlcNAc) modification on Ser/Thr has also been observed in nuclear and cytosolic proteins, which is involved in metastasis of cancer cells.²⁰⁻²¹

Among the various O-glycosylation types, mucin-type O-glycosylation that contains an α -N-acetylgalactosamine (GalNAc) core is most frequently identified. For this group of glycans, eight distinct core structures have been discovered, and they are illustrated in Figure 2. It is noteworthy that some of these structures (e.g. Core 1 and Core 8) share the same glycan compositions but differ in the linkage positions between adjacent monosaccharide units.^{19,}
²²Additional carbohydrates are often added to these core structures through enzyme-controlled reactions to generate O-glycans of diverse branching and elongation patterns.

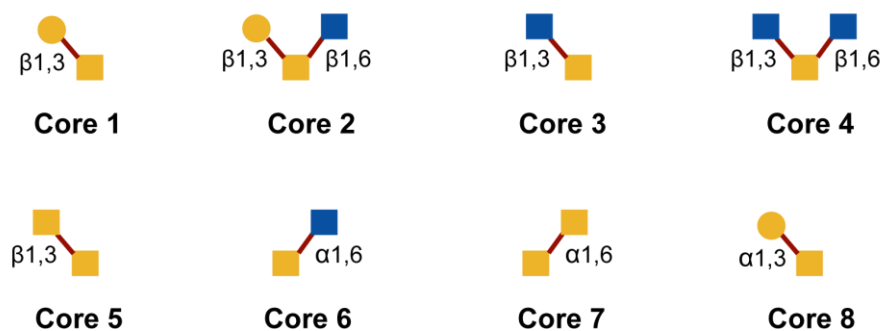


Figure 2. The eight core structures of mucin-type O-linked glycans. The linkage patterns between monosaccharide units are also shown in the figure. Glycan symbols are, blue square: N-acetylglucosamine (GlcNAc), yellow square: N-acetylgalactosamine (GalNAc), yellow circle: galactose (Gal).

1.1.4 Site-Specific Glycosylation Analysis

Protein glycosylation can be studied in two approaches: glycan-based analysis and glycopeptide-based analysis. The former approach involves the detachment of glycans from proteins and focuses on characterizing the glycans. Several excellent reviews are available that summarize methods and techniques used in glycomics analysis.²²⁻²⁶ Compared to the glycan-based approach, glycopeptide analysis has the advantage that the glycans remain attached to the peptide

sequences, such that glycosylation profiles can be directly related to the attachment sites. A schematic diagram illustrating the workflow of glycopeptide analysis is shown in Figure 3. The glycoproteins of interest are isolated from biological matrices, and they are subjected to proteolysis to generate mixtures of glycopeptides and peptides. The glycopeptides can be further enriched or separated, followed by detection using mass spectrometry (MS). The resulting MS data are submitted to software programs for data analysis and glycopeptide identification. A wide range of avenues have been employed to improve this workflow, which lead to further development and innovation in biological sample pretreatment, separation, and MS analysis of glycopeptides. In this dissertation, we describe some of the latest advances in these areas and discuss the relevant applications based on these new methods and tools.

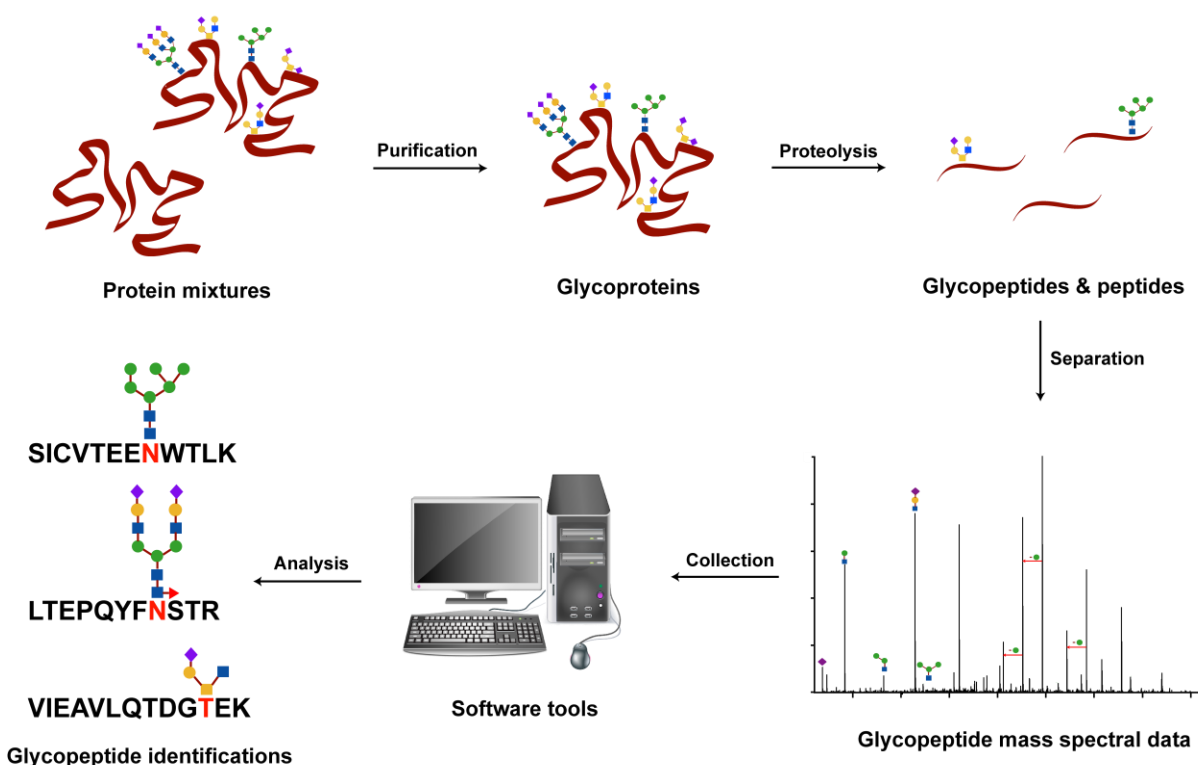


Figure 3. A general diagram of the glycopeptide-based analysis workflow by mass spectrometry.

1.2. Glycoprotein Purification and Digestion

1.2.1 Glycoprotein/Glycopeptide Enrichment

Because a variety of glycans can be present on the same glycosylation site (microheterogeneity), different protein glycoforms exist for each protein; consequently, the resulting glycopeptides are in lower abundance compared to non-glycosylated peptides. Therefore, sample purification is vital when complex biological samples (e.g. serum) containing a mixture of proteins and glycoproteins are analyzed. Glycoprotein purification can be performed at either the crude protein mixture level or at the glycopeptide level after proteolysis of glycoproteins.²⁷

Different types of lectins are heavily used to enrich glycoproteins that carry specific carbohydrate motifs through affinity binding. A list of lectins commonly utilized for glycosylation analysis, along with their specificities, is summarized in Reference 24.²⁸ In many applications, lectin columns are employed to extract glycoproteins with targeted glycan types from the matrix. For example, three lectins that bind to high-mannose glycans, concanavalin A (Con A), snowdrop lectin and lentil lectin, were used in parallel to enrich N- glycoproteins in tomato fruit pericarp, and the N-glycoproteome coverage was significantly increased due to the usage of multiple lectins.²⁹ An automated multi-lectin affinity chromatography (M-LAC) platform that consists of lectin mixtures of Con A, jacalin and wheat germ agglutinin (WGA), was used in conjunction with antibody columns that deplete albumin and IgG (major proteins in serum), to fractionate glycoproteins in pancreatic cancer fluid samples.³⁰⁻³² Lectin-based interactions can also be applied to glycopeptide enrichment, and Medzihradszky *et al.* used affinity columns packed with agarose-bound jacalin to extract mucin core-1 type O-linked glycopeptides from bovine serum.³³⁻³⁴ Lectin based strategies have also been useful in the identification of glycopeptide-based cancer biomarkers.³¹⁻³²

Apart from lectin-based methods that target carbohydrate structures, glycoproteins can also be isolated using antibodies that recognize epitopes on protein backbones. An important example is the analysis of transferrin, which is a glycoprotein that has altered glycoforms under normal and disease states in human blood plasma.³⁵ Heywood *et al.* used rabbit anti-transferrin antibody immobilized on magnetic beads to capture serum transferrin by immunoprecipitation, and the characterized glycosylation profiles were then employed to diagnose congenital disorders of glycosylation (CDG);³⁶ in another case, goat polyclonal antibody against transferrin was utilized to prepare immunoaffinity chromatography columns that isolated transferrin for the diagnosis of alcohol dependence.³⁷

Hydrophilic interaction liquid chromatography (HILIC) has gained more attention in the past few years as an effective tool to separate glycopeptides from peptides. Similar to normal phase liquid chromatography (NPLC), stationary phases with relatively high polarity, such as cellulose, sepharose, silica bonded with amine, cyanide and amide groups, have been used in HILIC;³⁸ on the other hand, the mobile phases in HILIC are similar to those utilized in reversed-phase liquid chromatography (RPLC), which are moderately polar solvents (e.g. a mixture of acetonitrile and water).³⁹ In HILIC, glycopeptides are retained by the polar stationary phase via hydrophilic interactions with the glycan portion while peptides are easily washed off. Elution of glycopeptides can be achieved by increasing the content of water in the mobile phase. A recent application uses amine functionalized magnetic nanoparticles with greater surface area to selectively enrich sialylated glycopeptides, based on hydrophilic interactions and weak anion exchange principles.⁴⁰ Zwitterionic-HILIC (ZIC-HILIC), a variation of HILIC that uses zwitterions as stationary phases, is found to have better separation efficiency for glycopeptides because additional ion-pairing effects exist in the separation process.⁴¹⁻⁴² An excellent review detailing the usage of HILIC in glycopeptide analysis is published elsewhere.⁴³

Glycoproteins and glycopeptides can also be extracted from matrices through covalent bond formation that targets on the *cis*-diol groups in carbohydrates. Boronic acid reacts with *cis*-diol-containing glycans to form boronic esters, and the reaction can be reversed at acidic pH conditions.^{38,44} Consequently, materials functionalized with boronic acid groups have been used in the enrichment of glycopeptides from human serum.⁴⁵ Hydrazide chemistry is another technique that utilizes covalent linkages to isolate glycoproteins, where the glycan *cis*-diols are oxidized to aldehydes and are bonded to the hydrazide groups on solid particles.⁴⁶ Nonetheless, the procedure is irreversible and the glycan portion is subsequently cleaved using glycosidases or chemical derivatization such that the glycosylation sites can be determined. The site-specific glycopeptide information, though, is not available using this approach.⁴⁶⁻⁴⁷

1.2.2 Glycoprotein Digestion

Proteolysis of glycoproteins is a pivotal step in sample preparation during which glycoproteins are cleaved into peptides and glycopeptides. Trypsin is typically used in glycopeptide analysis,⁴⁸ and other endoproteinases including Lys-C and Glu-C can also be employed to generate glycopeptides of appropriate lengths.⁴⁹⁻⁵⁰ Recently, immobilized enzyme reactor (IMER), a technique where enzymes are immobilized on solid supports as opposed to conventional in-solution digestion, has gained more applications in glycoproteomics because of the improved digestion efficiency, reduced digestion time and suitability for automation.⁵¹ Wang *et al.* encapsulated trypsin into a hydrophilic polyacrylate network by acryloylation and *in situ* polymerization.⁵² After reduction and alkylation, glycoprotein solution was pumped through the trypsin-modified polyacrylate network at a flow rate of 1 $\mu\text{L}/\text{min}$, and the digested glycopeptides were eluted off the reactor and detected. Using the IMER, the digestion time was reduced from 16

hours to several minutes, and the reactor could be re-cycled for 10 times with good reproducibility.⁵²

Glycoprotein digestion can also be performed after protein separation by gel electrophoresis, with targeted glycoprotein bands excised from the gel band.⁵³ Lebrilla and coworkers developed a method to separate protein mixtures in crude bovine milk and human serum by one-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (1-D SDS-PAGE), followed by in-gel nonspecific proteolysis of targeted glycoproteins (lactoferrin and transferrin).⁵⁴ Proteases with broad specificities, including pronase and pepsin, were used to yield glycopeptides of smaller sizes for site-specific glycosylation analysis.⁵⁴

Complementary to proteolysis, glycosidase digestion may be a viable route to facilitate glycopeptide-based analysis. Two glycosidases, endoglycosidase H (Endo H) and endoglycosidase F3 (Endo F3), were employed to partially deglycosylate HIV-1 envelope glycoproteins;⁵⁵ both enzymes cleave the glycosidic bond connecting the two GlcNAc residues in the conserved core, thus leaving the N-glycosylation site linked to one GlcNAc residue. However, Endo H selectively cleaves high-mannose and hybrid-type glycans, while Endo F3 cleaves only complex glycans.⁵⁶ By performing Endo H and Endo F3 digestion on separate aliquots of HIV samples followed by trypsin digestion, the glycan heterogeneity of the resulting glycopeptides were effectively reduced so that glycopeptides of poor ionization efficiency and low abundance were detected. Furthermore, di-glycosylated peptides, where two N-glycans attach to the same peptide, were well characterized using this approach because partial deglycosylation leads to glycopeptides containing one site occupied with a single GlcNAc and the other with an intact glycan that does not match the glycosidase specificity, thus greatly simplifying the MS analysis.⁵⁵

1.3. Analytical Separation of Glycopeptides

In order to facilitate the detection of glycopeptides by MS, a separation method is necessary to enrich glycopeptides from peptides that co-exist in the digested glycoprotein samples. Due to its capacity to resolve complex proteomics samples, high compatibility with MS detection, and the readily available automation system, liquid chromatography (LC) is the separation method of choice that is coupled to mass spectrometers for analyzing glycopeptides.²⁵ Offline LC fractionation of glycopeptides is one feasible approach to achieve separation to a certain extent, while online LC-MS analysis is more routinely adopted in glycoproteomics.

Reversed-phase LC columns with C₁₈ or C₈ bonded stationary phases are often used in glycopeptide separation, and the retention of glycopeptides is mainly determined by the peptide portion: therefore, glycoforms of the same peptide sequence have similar retention times.⁵⁷ When a glycosylation site is only partially occupied, the non-glycosylated form of the peptide elutes off earlier in RPLC than the glycopeptides, due to the hydrophilicity of the glycan portion. More recently, RPLC has been used in conjunction with other techniques that are chromatographically orthogonal, such as HILIC and ion-exchange chromatography, to accomplish improved separation and identification of glycopeptides.⁵⁸ Parker *et al.* combined ZIC-HILIC with C₁₈-based nanoflow RPLC-MS/MS analysis to enrich, separate and characterize intact N-glycopeptides from membrane-associated proteins, and 863 unique glycopeptides were confidently identified from 161 rat brain glycoproteins using their methods.⁵⁹

Due to a high polarizability of the surface, porous graphitized carbon (PGC) has the special ability to retain polar compounds, and PGC columns are extensively used in glycan and glycopeptide analysis.⁶⁰ Because the charge or dipole of the analyte induces polarization on the graphite surface, the spatial orientation of the functional groups of the analyte would affect the retention time. Consequently, structural isomers can be separated by PGC chromatography. As an

application in analyzing glycopeptides, PGC was packed into a microfluidic LC-chip by Nwosu *et al.*, who used a PGC chip to separate glycopeptides carrying heterogeneous glycan moieties that were not well resolved on C₈ or C₁₈ columns.⁶¹ A total of 233 distinct glycopeptides representing 18 glycosylation sites were identified in a single mixture. Figure 4 illustrates the chromatogram where glycopeptides were separated by the PGC chip.⁶¹ Other applications of PGC in glycopeptide separation are summarized elsewhere.^{54, 62}

Monolithic columns are a relatively new tool in separation science; they contain polymeric stationary phases that are continuous in structure. These columns can be conveniently prepared by *in-situ* polymerization reactions, and they possess excellent chromatographic properties including low backpressure, efficient mass transfer, and large surface area.⁶³ In an application where glycopeptides were separated using monolithic columns, narrow bore (10 μ m i.d.) porous layer open tubular (PLOT) columns were developed by Karger and coworkers through co-polymerization of styrene and divinylbenzene.⁶⁴⁻⁶⁶ At an ultralow flow rate of 20 nL/min, low abundant glycopeptides at about 100 amol level were identified and quantified.⁶⁶ Other types of monomers and reactants for co-polymerization have also been utilized to prepare appropriate monolithic materials that separate glycopeptides in HILIC mode⁶⁷ and in affinity chromatography.⁴⁴

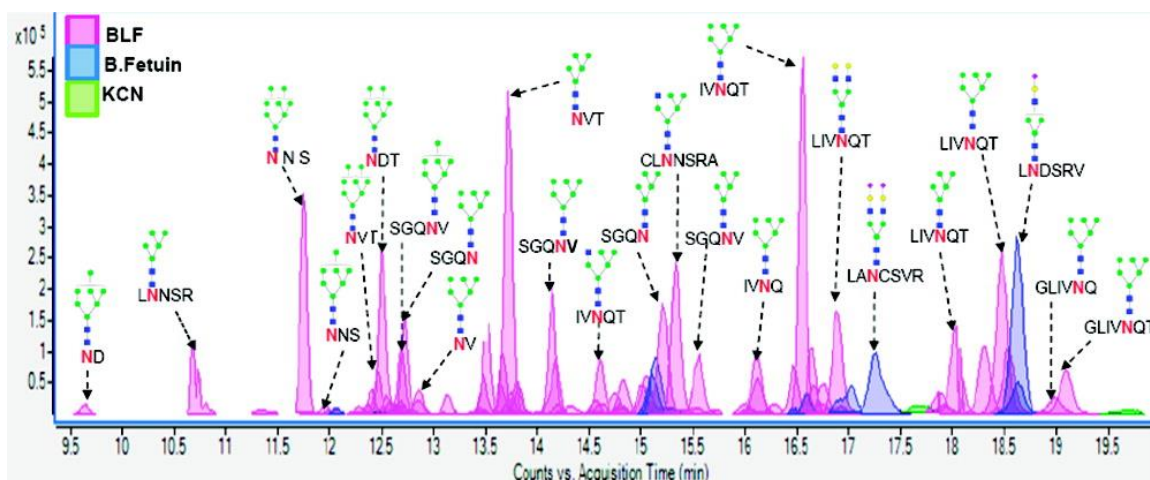


Figure 4. Extracted ion chromatograms of glycopeptides separated by a PGC LC-chip. Glycopeptides were generated from a mixture of glycoproteins containing bovine lactoferrin, kappa casein, and bovine fetuin. Reprinted with permission from Reference 59. Copyright 2011 American Chemical Society.

1.4. Mass Spectrometric Analysis of Glycopeptides

1.4.1 ESI and MALDI Mass Analysis

Electrospray ionization (ESI) is the most commonly used ionization technique in MS analysis of glycopeptides; it is very soft such that intact glycopeptide ions are formed and guided into the mass analyzer without structural decomposition or loss of labile groups.^{23, 68} In addition, multiply charged glycopeptide ions are typically formed through ESI, which is advantageous in some cases because the m/z values of large glycopeptides would be lowered into ranges that are measurable by appropriate mass analyzers.⁶⁹ Moreover, it affords the user the opportunity to fragment relevant glycopeptide ions by electron transfer dissociation (ETD).^{7, 55}

Apart from ESI, matrix-assisted laser desorption ionization (MALDI) is another highly effective tool to ionize glycopeptides through which singly charged glycopeptide ions are produced predominantly.⁷⁰ Therefore, the complexity of the MALDI-MS spectrum is greatly reduced compared to ESI-MS, and glycopeptides would have higher signal intensity because they would not split into different charge states.⁷¹ One example showing the use of MALDI is the work

of Watanabe *et al.* These authors used a new liquid matrix (3-aminoquinoline/ α -cyano-4-hydroxycinnamic acid) in MALDI-ion trap-TOF MS analysis of a glycoprotein, human epidermal growth factor 2 (HER2).⁷² Compared to MALDI experiments using conventional liquid matrices such as 2,5-dihydroxybenzoic acid (DHB), five novel glycopeptides carrying fucosylated complex-type and high-mannose glycans were identified by their method.⁷² MALDI-MS can also be adopted in high throughput glycosylation analysis of protein biomarkers. Wuhrer and coworkers recently performed a large scale N-glycosylation profiling of human immunoglobulin G (IgG) using a MALDI-TOF/TOF mass spectrometer in the negative mode. Plasma samples were collected for over 1709 individuals, and a clear trend of decreased galactosylation and sialylation was observed for IgG glycopeptides from individuals of increasing ages.⁷³

1.4.2 Tandem MS of Glycopeptides

Although glycopeptides with similar m/z values may be differentiated by high resolution mass spectrometers such as an Orbitrap (with mass error below 10 ppm), unambiguous assignment is relatively difficult to make based solely on mass analysis, especially for isobaric glycopeptides that differ in glycan linkages or peptide sequences.⁶⁸ Since single stage MS data only provides the m/z (thus the mass) of a glycopeptide, tandem MS (also called MS/MS) analysis is required to isolate and fragment the precursor ion from which detailed structural information can be obtained.⁷⁴⁻⁷⁵

1.4.2.1 Collisional Dissociation Methods

Collision-induced dissociation (CID), a collision-based activation method available in most commercial instruments including ion trap,⁶⁶ triple quadrupole,⁷⁶ quadrupole time-of-flight

(Q-TOF)⁷⁷ and other hybrids,⁷⁸⁻⁸⁰ is routinely used in MS/MS study of glycopeptides. Figure 5 presents a CID spectrum of a sialylated complex-type N-glycopeptide collected using a linear ion trap mass spectrometer. The data exhibits CID's general feature that peaks resulting from glycosidic bond cleavages dominate the spectrum, so that the glycan composition and connectivity of monosaccharide units are readily determined. Moreover, distinct Y1 ion peaks corresponding to peptide+GlcNAc are usually of high intensity in the CID data (as shown in Fig. 5), and they are used to infer possible peptide portion of the glycopeptide.⁸¹ Nevertheless, little or no peptide sequence information is available when glycopeptides are subjected to CID. Recent efforts to increase the peptide backbone dissociation of glycopeptides in CID are published: in these works, the collisional energy was modulated synergistically during the data acquisition period so that the energy-resolved CID spectrum would contain both glycan and peptide fragmentation.⁸²⁻⁸³ However, the method does not yet provide consistent backbone fragmentation across all different glycopeptides undergoing CID.

A specialized CID technique called higher-energy collisional dissociation (HCD), has been increasingly used in glycopeptide analysis by users of the hybrid ion trap-Orbitrap MS.⁸⁴⁻⁸⁶ The HCD fragmentation of glycopeptides is highly similar to that of beam-type CID, in which glycan oxonium ions of specific m/z values (e.g. HexNAc, m/z 204.2; HexNAc1Hex1, m/z 366.1) are prominent and can be used as diagnostic ions to indicate the presence of glycopeptides.⁸⁴ By combining HCD fragmentation with high mass accuracy Orbitrap detection, 88 previously uncharacterized glycopeptides were identified by Raftery *et al.* from 666 precursor ions in the analyses of hen egg glycoprotein digests.⁸⁵

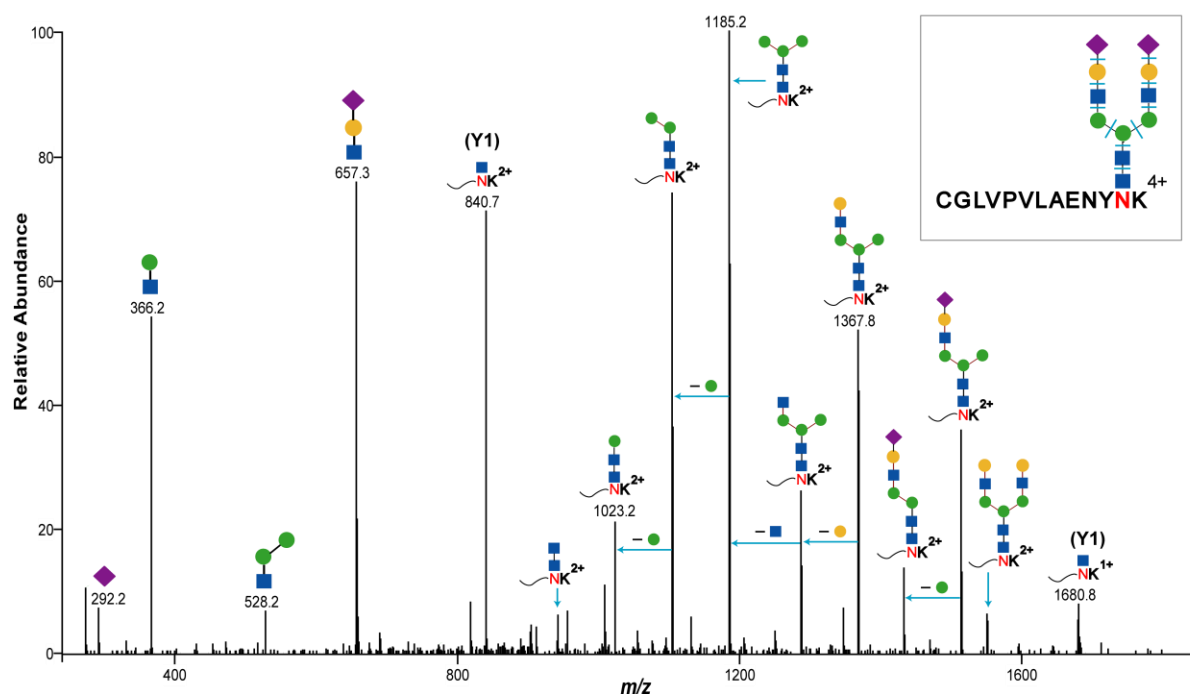


Figure 5. CID-MS/MS data of a biantennary complex N-linked glycopeptide from trypsin digested transferrin. The peptide sequence and the glycan structure are shown in the figure. Based on these data, the glycan composition, as well as the peptide mass, can be readily determined.

1.4.2.2 Electron-Based Fragmentation

Two electron-based fragmentation techniques, electron capture dissociation (ECD)⁸⁷ and electron transfer dissociation (ETD),⁸⁸ are becoming mainstream tools in glycoproteomics. In ECD and ETD, significant peptide backbone dissociation occurs on the glycopeptide while the glycan part remains intact, which is highly complementary to the dissociation pathway in CID. As a result, glycopeptides are oftentimes characterized by a combination of CID and ECD/ETD approaches,⁸⁹⁻⁹¹ and glycopeptide sequences, as well as the glycosylation sites, can be determined based on ECD/ETD data.

ECD has traditionally been achieved on FT-ICR mass spectrometers.⁹²⁻⁹³ In a recent report though, ECD fragmentation of sialylated N-glycopeptides was performed on a radio frequency-quadrupole ion trap.⁴⁹ By increasing the electron energy from 0.2 eV to over 4 eV, the

glycopeptide's sequence coverage was improved considerably.⁴⁹ In another application, human IgA, an antibody that plays a critical role in the mucosal immune system, was interrogated at its 9 potential O-glycosylation sites in the hinge region by high resolution MS and ECD tandem MS analysis.⁹⁴ O-linked glycopeptide structural isomers that differ by the glycosylation site (positional isomers) and by the glycan chain (glycan isomers) were fully differentiated using ECD, and a series of hinge region O-glycopeptides were identified from both the IgA1 myeloma protein and normal serum IgA1.⁹⁴

ETD is most frequently used to generate ECD-like fragmentation on ion trap instruments,⁹⁵ and it has been observed that the efficiency of glycopeptide dissociation in ETD increases when the glycopeptide ion carries more charges.⁹⁶ An example is given in Figure 6, where asialofetuin glycopeptides of varying charge states were fragmented by ETD. The c and z ions that result from peptide backbone cleavages are much more pronounced when the glycopeptide precursor is at 6+, whereas no fragment ions except the charge-reduced precursors are found for the same glycopeptide at 4+ charge state.⁹⁷ To enhance the fragmentation efficiency in ETD, m-nitrobenzyl alcohol (m-NBA), a supercharging reagent, was added to the LC eluate after column separation in order to increase the average charge state of glycopeptides.⁹⁸ As a result, highly charged N- and O-linked glycopeptides from recombinant monoclonal antibody trastuzumab and erythropoietin were produced in ESI-MS, and ETD-MS/MS analyses of these ions revealed both glycosylation sites and full sequence coverage of the analytes.⁹⁸

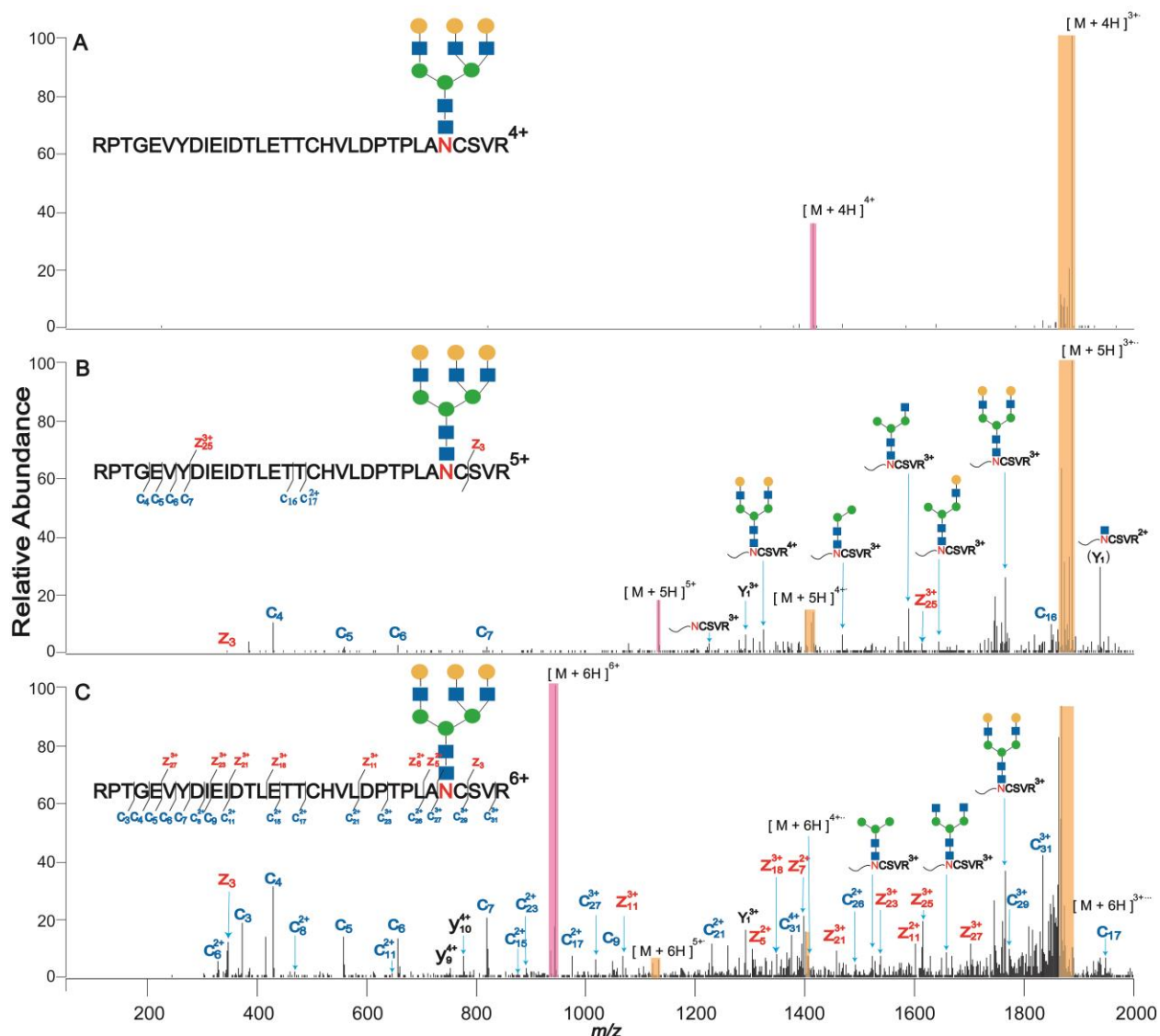


Figure 6. ETD data of a glycopeptide carrying a triantennary complex N-glycan at different charge states: (A) 4+, (B) 5+, (C) 6+. The precursor ion peak and peaks resulting from charge-reduced species are marked with color bars. Glycopeptide backbone fragments (c and z ions), along with some glycan dissociation products, are present in the figure. Reprinted with permission from Reference 97. Copyright 2013 American Chemical Society.

1.4.2.3 Photon-Induced Dissociation

Glycopeptide ions can also be activated via photon irradiation, during which the internal energy of the precursor is increased by absorbing photons to induce fragmentation. Infrared multiphoton dissociation (IRMPD) is a commonly used technique of this type, and it uses CO₂ lasers to irradiate infrared photons at μm wavelengths.^{7, 99} Glycopeptide fragmentation in IRMPD

is similar to what is observed in CID, with glycosidic bonds being preferentially cleaved as opposed to peptide backbone dissociation.¹⁰⁰ Another promising tool that relies on ion-photon interactions is ultraviolet photodissociation (UVPD),¹⁰¹ and it has recently been applied to glycopeptide analysis.¹⁰²⁻¹⁰³ Brodbelt and coworkers employed a 193 nm laser to perform UVPD on deprotonated O-linked glycopeptide anions from bovine kappa casein digest, and a representative glycopeptide UVPD spectrum is shown in Figure 7. One hundred percent sequence coverage is achieved, as demonstrated by an extensive a and x ion series. Moreover, the intact O-glycan structure is retained on all the peptide fragments (a_5 - a_9 and x_6 - x_9) that contain the glycosylated Thr, hence the glycosylation site is confidently assigned.¹⁰² Significant glycosidic cleavages are also identified in the same UVPD data, from which the glycan composition can be determined. The concurrent sequencing ability of both the glycan and peptide moieties renders UVPD a promising tandem MS method for elucidating glycopeptide structures.

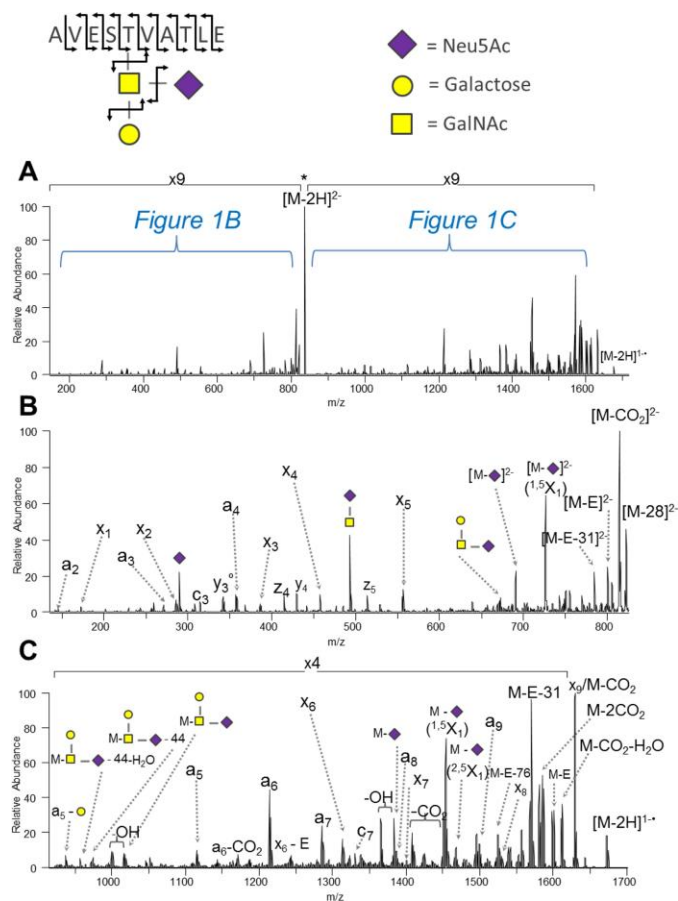


Figure 7. UVPD data of a doubly deprotonated O-glycopeptide anion from kappa casein, whose structure is depicted in the figure. (A) is the entire spectrum, while (B) and (C) are expansions of the low and high m/z regions in (A), respectively. Reprinted with permission from Reference 102. Copyright 2013 American Chemical Society.

1.4.3 Ion Mobility Spectrometry

Ion mobility spectrometry (IMS) separates ionized molecules on the basis of their mobility in a carrier gas, where mobility is determined by factors including size and shape of the analytes.¹⁰⁴ As a result, IMS adds an orthogonal separation dimension to the capability of mass spectrometers that distinguish ions by m/z values. A few studies have been reported on the usage of IMS to analyze glycopeptides. Damen *et al.* performed tandem MS analysis of N-linked glycopeptides using quadrupole-ion mobility-TOF MS, where the glycopeptide was subjected to CID in the first collision cell, followed by separation of the resulting fragment ions using IMS so that fragments

were guided into a second collision cell at different times. The Y1 ion (peptide+GlcNAc) was further fragmented by activating high voltage of the second collision cell only when Y1 ion passed through the ion mobility drift tube to enter into the second cell. In this way the glycopeptide sequence could be probed by interpreting the tandem MS data collected at specific ion mobility drift times, which is similar to an MS³ experiment performed on an ion trap instrument.¹⁰⁵ The plot of the ion mobility data showing the separation of the Y1 ion is demonstrated in Figure 8. In another application, two O-linked mucin glycopeptides that have identical sequences but differing glycosylation sites (positional isomers) were successfully separated by high field asymmetric wave ion mobility spectrometry (FAIMS), which shows the great potential of this technique in differentiating structural isomers of glycopeptides.¹⁰⁶

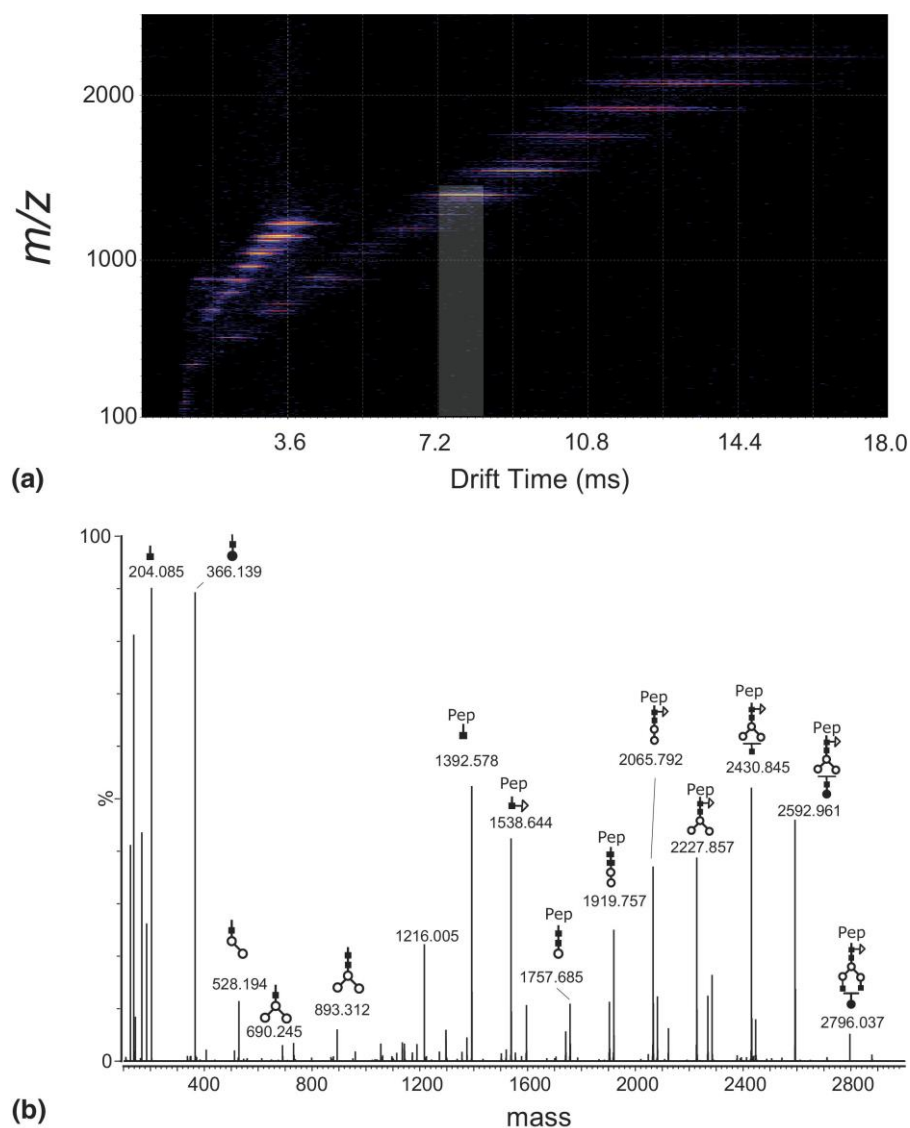


Figure 8. Glycopeptide analysis by quadrupole-ion mobility-TOF MS. (a) Two-dimensional graph of ion drift time versus m/z values for fragmentation ions resulting from CID of a glycopeptide in the first collision cell, followed by ion mobility separation. The white bar area indicates the Y1 ion (m/z 1392.58) with a drift time of 7.2-8.3 ms, and this ion was further fragmented in the second collision cell to generate MS3-like spectrum (data not shown). (b) MS/MS data of the same glycopeptide precursor in (a), but with no ion mobility separation or secondary fragmentation. Reprinted with permission from Reference 105. Copyright 2009 American Society for Mass Spectrometry.

1.5 Automated Glycoproteomics

With more and more software being developed that can interpret glycopeptide mass spectral data, automated glycoproteomics analysis is rapidly becoming a method of choice in

many glycopeptide-centric applications.¹⁰⁷ It is particularly useful when a large amount of data need to be analyzed; by choosing the appropriate software program, glycopeptide assignment can be made in a high-throughput fashion, and the false discovery rate of the identifications can also be determined.¹⁰⁸

A majority of the currently available bioinformatics tools utilize both MS and MS/MS data to decipher glycopeptide compositions, with CID data being investigated most.¹⁰⁹⁻¹¹³ A program called GlycoPeptideSearch (GPS) was recently developed that assigns N-linked glycopeptides based on CID spectra.^{77, 111} Fragment ions that represent the intact peptide plus monosaccharides of the N-glycan core (GlcNAc, GlcNAc-GlcNAc, GlcNAc-GlcNAc-Man) are identified by GPS so that the mass of the peptide portion can be determined; the glycan mass is thus calculated by subtracting the peptide mass from the precursor ion mass. GPS subsequently searches these two masses through the protein sequence database provided by the user and a glycan database called GlycomeDB respectively, to compute possible peptide-glycan pairs. Lastly, the isotope cluster profile in the MS data is compared to the theoretical isotope distribution of each peptide-glycan pair in order to determine the quality of each match.¹¹¹

A different approach was taken by Woodin *et al.*, who developed an online software tool called GlycoPep Grader (GPG) that assigns the best-matching glycopeptide composition to each CID spectrum.¹¹² GPG works in conjunction with a separate program such as GlycoMod¹¹⁴ that generates a list of possible glycopeptide candidates based on high resolution MS data. By calculating and searching for potential peaks that denote glycan dissociations of the precursor ion, GPG scores every glycopeptide candidate based on the CID data and determines the most likely glycopeptide composition for every spectrum. A low false discovery rate was achieved by GPG when it was utilized to analyze glycoproteomics data from different proteins.¹¹²

A new software named GlycoPeptide Finder (GP Finder) was introduced in 2013 that can accommodate a nonspecific protease digestion workflow.¹¹³ Peptide tags of varied lengths and sequences that contain potential glycosylation sites are paired with glycans from glycan libraries to comprise possible glycopeptide candidates, which are subsequently scored against the tandem MS data to determine the best candidate. The confidence of the assignments are improved by self-consistency in that each glycan would be detected to be present on multiple peptides of different lengths that encompass the same glycosylation site.¹¹³ However, data analysis is still quite complicated when nonspecific enzymes are used in proteolysis because the resulting peptide tag is highly variable and sometimes too short for site-specific identification.

In view of the great complementarity that ETD leads to, a few programs have included the functionality to analyze ETD data of glycopeptides.^{97, 115-118} This functionality is particularly useful in O-linked glycopeptide analysis, because no consensus sequence exists to predict the potential O-glycosylation site, which is also hard to determine by CID data.¹¹⁹⁻¹²¹ Important work has been done by Darula *et al.* that enabled the Protein Prospector software to identify O-glycopeptides carrying GalNAc-Gal₀₋₁-Neu5Ac₀₋₁ by scoring the ETD spectra.¹¹⁵⁻¹¹⁶ The software allows merging of CID and ETD search results, and comparison of different modification site assignments can be made by the user for site localization. To automate the glycopeptide identification workflow by ETD, we studied the fragmentation patterns of N- and O-linked glycopeptides from model glycoproteins and developed novel algorithms to consider multiple ion series (c, z and y ions) of putative glycopeptide candidates separately.^{97, 117} By implementing the algorithms into standalone software programs that handle glycopeptide ETD data, the correct glycopeptide compositions and site assignments were made in a high throughput and high accuracy manner.

Recently, a key trend in the development of versatile bioinformatics tools for glycopeptides is the incorporation of different types of tandem MS data into the analysis workflow. For instance, Tang *et al.* updated the GlypID program so that CID, HCD and high resolution MS data can be incorporated together into improved characterization of N-linked glycopeptides.^{109, 122} The HCD data provided the m/z values of glycan oxonium ions and the Y1 ion (peptide+GlcNAc) detected by the Orbitrap with high mass accuracy, and these fragments were compared with the peaks found in the CID data to confirm the glycopeptide's glycan composition. GlypID can also interpret MS³ data resulting from the fragmentation of the Y1 ion to identify the glycopeptide sequence based on b and y ions.¹²² Another software tool with wide functionality is GlycoFragwork, which has different scoring algorithms for CID, HCD and ETD fragmentation of glycopeptides.¹¹⁸ The HCD scoring was used to filter non-glycopeptide ions by examining the presence of diagnostic oxonium ions, while CID and ETD analysis were employed to determine the glycan sequence (monosaccharide composition and topology) and the peptide sequence, respectively. By utilizing GlycoFragwork, over 100 glycopeptides of 53 distinct glycosylation sites across 33 proteins were identified with high confidence from complex human serum samples.¹¹⁸ Nonetheless, the advancement of glycopeptide-based software tools still falls short compared to the development of new analytical techniques that are used in glycoproteomics. It is certain that more attention and effort need to be put into the data analysis part in the study of glycoproteins, thus extending new possibilities in large-scale glycoproteome analysis.

1.6 Conclusions

Investigating protein glycosylation has profound significance in uncovering the many roles that carbohydrates play in numerous cellular activities, and the knowledge gained will provide insights into understanding disease pathogenesis¹²³⁻¹²⁴ as well as developing new

medicine.¹²⁵⁻¹²⁶ However, unlike the biosynthesis of DNA, RNA and proteins, glycans are not synthesized through template-driven pathways, and complex branching and linkage patterns have been observed on these carbohydrate structures. Moreover, in glycopeptide analysis, the glycan information needs to be obtained in the context of specific glycosylation sites, which can provide a significant challenge to the investigator. Fortunately, recent advances in multiple aspects of glycoproteomics, including efficient glycoprotein digestion, facile LC separation of structurally similar glycopeptides, MS detection with high mass accuracy and varied fragmentation modes, software tools for high throughput data analysis, help to address some of the important questions in glycobiology that were difficult to tackle with before. The development of new technology in glycopeptide analysis will continue, and the capabilities of different analytical strategies (glycomics, proteomics and glycoproteomics) can be combined in future study of glycosylated proteins.

1.7 Summary of Subsequent Chapters

Chapter 2 describes the fragmentation patterns of different N-linked glycopeptide species in electron transfer dissociation (ETD), and the fragmentation rules are summarized to build a software tool called GlycoPep Detector (GPD). This is the first report in the glycoproteomics field where a computer program is developed to interpret the ETD data, such that glycopeptide assignments can be made in an automated fashion. In this study, unique fragmentation features are discovered for glycopeptides of varied sequences and glycan compositions, in that one fragment ion series (c- or z-ions) dominate the ETD spectrum, with the other ion series completely unrecorded, due to the existence of the glycan sidechain. Based on the distinct spectral features, an effective algorithm is designed to incorporate the fragmentation rules of N-glycopeptides in ETD, and the developed program, GPD, assigns the highest score to the correct glycopeptide in every test, with zero false positive identification.

Chapter 3 explores the fragmentation behaviors of the other type of glycopeptides, O-linked glycopeptides, in ETD mass spectrometry. A software named GlycoPep Scorer (GPS) is developed that can make automated O-glycopeptide assignments in combination with the ETD fragmentation method. The result of the automated analysis using GPS is directly compared with that of a publicly available program, ProteinProspector, in analyzing ETD datasets of various O-linked glycopeptides. These O-glycopeptides originate from different biological sources including bovine fetuin, erythropoietin, mucin and the HIV envelope glycoprotein gp120. It is demonstrated that GPS leads to a larger score difference between the correct glycopeptide candidate and the best-scoring incorrect candidate, thus facilitating O-glycopeptide identification with improved accuracy. Furthermore, fewer false positive hits are found using the GPS tool compared to ProteinProspector.

Chapter 4 details the usage of a software program called GlycoPep Evaluator (GPE), in determining the false discovery rate (FDR) of automated glycopeptide assignments. Different from the conventional approach where a decoy database is created based on the target protein sequence database, no decoy database is utilized in our approach. Instead, as many as 20 decoy glycopeptides are created *de novo* using the GPE program. Subsequently, the decoys, along with target glycopeptides, are scored against the ETD data, and the FDRs in glycopeptide identifications can be determined accurately, based on the number of decoy matches and the target-to-decoy ratio (e.g. 1:20 in our case). This method proves to be particularly useful in analyzing small data sets when a limited number of glycopeptide ETD spectra are available, which is often the case when a single glycoprotein is characterized.

Chapter 5 introduces a novel quantitation method for determining the N-glycosylation site occupancy of glycoproteins. The site occupancy indicates the extent of glycosylation on a protein, and this value can change even for the same protein that is expressed under different cell lines. In order to measure the glycosylation site occupancy, the most commonly adopted approach uses protein N-glycosidase F (PNGase F) to detach the N-glycan from the protein, through which the glycosylated asparagine is converted to aspartic acid. After proteolysis, the glycosylation site occupancy can be determined by comparing the MS signal of the deglycosylated peptide against the non-glycosylated peptide. However, chemical deamidation of asparagine occurs spontaneously during sample preparation, and the non-glycosylated asparagine that undergoes chemical deamidation would be incorrectly assigned as deglycosylated asparagine, thus leading to inaccurate quantitation result. In contrast, in the newly developed method, the non-glycosylated peptide is quantified using an isotopically labeled internal standard, and the total glycoprotein concentration is determined using a second set of peptide standard. No PNGase F is used in the new method, such that the N-glycan stays intact on the glycosylation site.

It is demonstrated that the developed approach leads to accurate quantitation of the glycosylation site occupancy of a model glycoprotein, fetuin, without the adverse impact from chemical deamidation.

References

- (1) Cummings, R. D.; Pierce, J. M. *Chem. Biol.* **2014**, *21*, 1-15.
- (2) Ohtsubo, K.; Marth, J. D. *Cell.* **2006**, *126*, 855-867.
- (3) Aeby, M. *Biochim. Biophys. Acta-Mol. Cell Res.* **2013**, *1833*, 2430-2437.
- (4) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. *Scientific Reports.* **2011**, *1*, 1-5.
- (5) Haltiwanger, R. S.; Lowe, J. B. *Annu. Rev. Biochem.* **2004**, *73*, 491-537.
- (6) Novotny, M. V.; Alley, W. R. *Curr. Opin. Chem. Biol.* **2013**, *17*, 832-840.
- (7) Dodds, E. D. *Mass Spectrometry Reviews.* **2012**, *31*, 666-682.
- (8) Wolfert, M. A.; Boons, G. J. *Nat. Chem. Biol.* **2013**, *9*, 776-784.
- (9) Christiansen, M. N.; Chik, J.; Lee, L.; Anugraham, M.; Abrahams, J. L.; Packer, N. H. *Proteomics.* **2014**, *14*, 525-546.
- (10) Bailey, U. M.; Jamaluddin, M. F.; Schulz, B. L. *J. Proteome Res.* **2012**, *11*, 5376-5383.
- (11) Beck, A.; Sanglier-Cianferani, S.; Van Dorsselaer, A. *Anal. Chem.* **2012**, *84*, 4637-4646.
- (12) Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta-Gen. Subj.* **1999**, *1473*, 4-8.
- (13) Matsui, T.; Takita, E.; Sato, T.; Kinjo, S.; Aizawa, M.; Sugiura, Y.; Hamabata, T.; Sawada, K.; Kato, K. *Glycobiology.* **2011**, *21*, 994-999.
- (14) Jones, J.; Krag, S. S.; Betenbaugh, M. J. *Biochim. Biophys. Acta-Gen. Subj.* **2005**, *1726*, 121-137.
- (15) Petrescu, A. J.; Milac, A. L.; Petrescu, S. M.; Dwek, R. A.; Wormald, M. R. *Glycobiology.* **2004**, *14*, 103-114.
- (16) Kuster, B.; Mann, M. *Anal. Chem.* **1999**, *71*, 1431-1440.

- (17) Segu, Z. M.; Hussein, A.; Novotny, M. V.; Mechref, Y. *J. Proteome Res.* **2010**, *9*, 3598-3607.
- (18) Zauner, G.; Kozak, R. P.; Gardner, R. A.; Fernandes, D. L.; Deelder, A. M.; Wuhler, M. *Biol. Chem.* **2012**, *393*, 687-708.
- (19) Jensen, P. H.; Kolarich, D.; Packer, N. H. *Febs J.* **2010**, *277*, 81-94.
- (20) Yi, W.; Clark, P. M.; Mason, D. E.; Keenan, M. C.; Hill, C.; Goddard, W. A.; Peters, E. C.; Driggers, E. M.; Hsieh-Wilson, L. C. *Science.* **2012**, *337*, 975-980.
- (21) Zachara, N. E.; Molina, H.; Wong, K. Y.; Pandey, A.; Hart, G. W. *Amino Acids.* **2011**, *40*, 793-808.
- (22) Alley, W. R.; Novotny, M. V. *Annu. Rev. Anal. Chem.* **2013**, *6*, 237-265.
- (23) Leymarie, N.; Zaia, J. *Anal. Chem.* **2012**, *84*, 3040-3048.
- (24) Rakus, J. F.; Mahal, L. K. *Annu. Rev. Anal. Chem.* **2011**, *4*, 367-392.
- (25) Suzuki, S. *Anal. Sci.* **2013**, *29*, 1117-1128.
- (26) Kailemia, M. J.; Ruhaak, L. R.; Lebrilla, C. B.; Amster, I. J. *Anal. Chem.* **2014**, *86*, 196-212.
- (27) Ongay, S.; Boichenko, A.; Govorukhina, N.; Bischoff, R. *J. Sep. Sci.* **2012**, *35*, 2341-2372.
- (28) Alley, W. R.; Mann, B. F.; Novotny, M. V. *Chemical Reviews.* **2013**, *113*, 2668-2732.
- (29) Ruiz-May, E.; Hucko, S.; Howe, K. J.; Zhang, S.; Sherwood, R. W.; Thannhauser, T. W.; Rose, J. K. C. *Mol. Cell. Proteomics.* **2014**, *13*, 566-579.
- (30) Gbormittah, F. O.; Haab, B. B.; Partyka, K.; Garcia-Ott, C.; Hincapie, M.; Hancock, W. S. *J. Proteome Res.* **2014**, *13*, 289-299.
- (31) Kullolli, M.; Hancock, W. S.; Hincapie, M. *Anal. Chem.* **2010**, *82*, 115-120.
- (32) Plavina, T.; Wakshull, E.; Hancock, W. S.; Hincapie, M. *J. Proteome Res.* **2007**, *6*, 662-671.
- (33) Darula, Z.; Medzihradszky, K. F. *Mol. Cell. Proteomics.* **2009**, *8*, 2515-2526.

- (34) Darula, Z.; Sherman, J.; Medzihradszky, K. F. *Mol. Cell. Proteomics*. **2012**, *11*.
- (35) Golka, K.; Wiese, A. *J. Toxicol. Env. Health-Pt b-Crit. Rev.* **2004**, *7*, 319-337.
- (36) Heywood, W. E.; Mills, P.; Grunewald, S.; Worthington, V.; Jaeken, J.; Carreno, G.; Lemonde, H.; Clayton, P. T.; Mills, K. *J. Proteome Res.* **2013**, *12*, 3471-3479.
- (37) Barroso, A.; Gimenez, E.; Benavente, F.; Barbosa, J.; Sanz-Nebot, V. *Anal. Chim. Acta.* **2013**, *804*, 167-175.
- (38) Chen, C. C.; Su, W. C.; Huang, B. Y.; Chen, Y. J.; Tai, H. C.; Obena, R. P. *Analyst.* **2014**, *139*, 688-704.
- (39) Buszewski, B.; Noga, S. *Anal. Bioanal. Chem.* **2012**, *402*, 231-247.
- (40) Bodnar, E. D.; Perreault, H. *Anal. Chem.* **2013**, *85*, 10895-10903.
- (41) Di Palma, S.; Boersema, P. J.; Heck, A. J. R.; Mohammed, S. *Anal. Chem.* **2011**, *83*, 3440-3447.
- (42) Boersema, P. J.; Mohammed, S.; Heck, A. J. R. *Anal. Bioanal. Chem.* **2008**, *391*, 151-159.
- (43) Zauner, G.; Deelder, A. M.; Wuhrer, M. *Electrophoresis.* **2011**, *32*, 3456-3466.
- (44) Li, H. Y.; Liu, Z. *Trac-Trends Anal. Chem.* **2012**, *37*, 148-161.
- (45) Wang, Y. L.; Liu, M. B.; Xie, L. Q.; Fang, C. Y.; Xiong, H. M.; Lu, H. J. *Anal. Chem.* **2014**, *86*, 2057-2064.
- (46) Chen, R.; Jiang, X. N.; Sun, D. G.; Han, G. H.; Wang, F. J.; Ye, M. L.; Wang, L. M.; Zou, H. F. *J. Proteome Res.* **2009**, *8*, 651-661.
- (47) Klement, E.; Lipinszki, Z.; Kupihar, Z.; Udvardy, A.; Medzihradszky, K. F. *J. Proteome Res.* **2010**, *9*, 2200-2206.
- (48) Rebecchi, K. R.; Go, E. P.; Xu, L.; Woodin, C. L.; Mure, M.; Desaire, H. *Anal. Chem.* **2011**, *83*, 8484-8491.

- (49) Manri, N.; Satake, H.; Kaneko, A.; Hirabayashi, A.; Baba, T.; Sakamoto, T. *Anal. Chem.* **2013**, *85*, 2056-2063.
- (50) Gimenez, E.; Ramos-Hernan, R.; Benavente, F.; Barbosa, J.; Sanz-Nebot, V. *Anal. Chim. Acta.* **2012**, *709*, 81-90.
- (51) Ma, J. F.; Zhang, L. H.; Liang, Z.; Shan, Y. C.; Zhang, Y. K. *Trac-Trends Anal. Chem.* **2011**, *30*, 691-702.
- (52) Wang, C. F.; Gao, M. X.; Zhang, P.; Zhang, X. M. *Chromatographia.* **2014**, *77*, 413-418.
- (53) Ghosh, D.; Beavis, R. C.; Wilkins, J. A. *J. Proteome Res.* **2008**, *7*, 1572-1583.
- (54) Nwosu, C. C.; Huang, J. C.; Aldredge, D. L.; Strum, J. S.; Hua, S.; Seipert, R. R.; Lebrilla, C. B. *Anal. Chem.* **2013**, *85*, 956-963.
- (55) Go, E. P.; Hewawasam, G.; Liao, H. X.; Chen, H. Y.; Ping, L. H.; Anderson, J. A.; Hua, D. C.; Haynes, B. F.; Desaire, H. *J. Virol.* **2011**, *85*, 8270-8284.
- (56) Tarentino, A. L.; Plummer, T. H. *Methods Enzymol.* **1994**, *230*, 44-57.
- (57) Desaire, H. *Mol. Cell. Proteomics.* **2013**, *12*, 893-901.
- (58) Wang, H.; Wong, C. H.; Chin, A.; Taguchi, A.; Taylor, A.; Hanash, S.; Sekiya, S.; Takahashi, H.; Murase, M.; Kajihara, S.; Iwamoto, S.; Tanaka, K. *Nat. Protoc.* **2011**, *6*.
- (59) Parker, B. L.; Thaysen-Andersen, M.; Solis, N.; Scott, N. E.; Larsen, M. R.; Graham, M. E.; Packer, N. H.; Cordwell, S. J. *J. Proteome Res.* **2013**, *12*, 5791-5800.
- (60) West, C.; Elfakir, C.; Lafosse, M. *J. Chromatogr. A.* **2010**, *1217*, 3201-3216.
- (61) Nwosu, C. C.; Seipert, R. R.; Strum, J. S.; Hua, S. S.; An, H. J.; Zivkovic, A. M.; German, B. J.; Lebrilla, C. B. *J. Proteome Res.* **2011**, *10*, 2612-2624.
- (62) Hua, S.; Nwosu, C. C.; Strum, J. S.; Seipert, R. R.; An, H. J.; Zivkovic, A. M.; German, J. B.; Lebrilla, C. B. *Anal. Bioanal. Chem.* **2012**, *403*, 1291-1302.
- (63) Ou, J. J.; Lin, H.; Zhang, Z. B.; Huang, G.; Dong, J.; Zou, H. F. *Electrophoresis.* **2013**, *34*, 126-140.

- (64) Yue, G. H.; Luo, Q. Z.; Zhang, J.; Wu, S. L.; Karger, B. L. *Anal. Chem.* **2007**, *79*, 938-946.
- (65) Luo, Q. Z.; Rejtar, T.; Wu, S. L.; Karger, B. L. *J. Chromatogr. A.* **2009**, *1216*, 1223-1231.
- (66) Wang, D. D.; Hincapie, M.; Rejtar, T.; Karger, B. L. *Anal. Chem.* **2011**, *83*, 2029-2037.
- (67) Malerod, H.; Rogeberg, M.; Tanaka, N.; Greibrokk, T.; Lundanes, E. *J. Chromatogr. A.* **2013**, *1317*, 129-137.
- (68) Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. *J. Chromatogr. B.* **2007**, *849*, 115-128.
- (69) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. *J. Proteome Res.* **2007**, *6*, 3995-4005.
- (70) El-Aneed, A.; Cohen, A.; Banoub, J. *Appl. Spectrosc. Rev.* **2009**, *44*, 210-230.
- (71) Wuhrer, M.; de Boer, A. R.; Deelder, A. M. *Mass Spectrometry Reviews.* **2009**, *28*, 192-206.
- (72) Watanabe, M.; Terasawa, K.; Kaneshiro, K.; Uchimura, H.; Yamamoto, R.; Fukuyama, Y.; Shimizu, K.; Sato, T. A.; Tanaka, K. *Anal. Bioanal. Chem.* **2013**, *405*, 4289-4293.
- (73) Bakovic, M. P.; Selman, M. H. J.; Hoffmann, M.; Rudan, I.; Campbell, H.; Deelder, A. M.; Lauc, G.; Wuhrer, M. *J. Proteome Res.* **2013**, *12*, 821-831.
- (74) Seipert, R. R.; Dodds, E. D.; Clowers, B. H.; Beecroft, S. M.; German, J. B.; Lebrilla, C. B. *Anal. Chem.* **2008**, *80*, 3684-3692.
- (75) Pan, S.; Chen, R.; Aebersold, R.; Brentnall, T. A. *Mol. Cell. Proteomics.* **2011**, *10*, 1-14.
- (76) Huddleston, M. J.; Bean, M. F.; Carr, S. A. *Anal. Chem.* **1993**, *65*, 877-884.
- (77) Pompach, P.; Chandler, K. B.; Lan, R.; Edwards, N.; Goldman, R. *J. Proteome Res.* **2012**, 1728-1740.
- (78) Sandra, K.; Devreese, B.; Van Beeumen, J.; Stals, I.; Claeysens, M. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 413-423.
- (79) Whelan, S. A.; Lu, M.; He, J. B.; Yan, W. H.; Saxton, R. E.; Faull, K. F.; Whitelegge, J. P.; Chang, H. *J. Proteome Res.* **2009**, *8*, 4151-4160.

- (80) Zhu, Z.; Go, E.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1-6.
- (81) Dalpathado, D. S.; Desaire, H. *Analyst.* **2008**, *133*, 731-738.
- (82) Kolli, V.; Dodds, E. D. *Analyst.* **2014**, *139*, 2144-53.
- (83) Vekey, K.; Ozohanics, O.; Toth, E.; Jeko, A.; Revesz, A.; Krenyacz, J.; Drahos, L. *Int. J. Mass Spectrom.* **2013**, *345*, 71-79.
- (84) Singh, C.; Zampronio, C. G.; Creese, A. J.; Cooper, H. J. *J. Proteome Res.* **2012**, *11*, 4517-4525.
- (85) Hart-Smith, G.; Raftery, M. J. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 124-140.
- (86) Segu, Z. M.; Mechref, Y. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 1217-1225.
- (87) Bakhtiar, R.; Guan, Z. Q. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 1-8.
- (88) Kim, M. S.; Pandey, A. *Proteomics.* **2012**, *12*, 530-542.
- (89) Yin, X. K.; Bern, M.; Xing, Q. R.; Ho, J.; Viner, R.; Mayr, M. *Mol. Cell. Proteomics.* **2013**, *12*, 956-978.
- (90) Mechref, Y. *Current protocols in protein science* **2012**, *68*, 1-11.
- (91) Halim, A.; Ruetschi, U.; Larson, G.; Nilsson, J. *J. Proteome Res.* **2013**, *12*, 573-584.
- (92) Cooper, H. J.; Hakansson, K.; Marshall, A. G. *Mass Spectrometry Reviews.* **2005**, *24*, 201-222.
- (93) Fornelli, L.; Schmid, A. W.; Grasso, L.; Vogel, H.; Tsybin, Y. O. *Chem.-Eur. J.* **2011**, *17*, 486-497.
- (94) Takahashi, K.; Smith, A. D.; Poulsen, K.; Kilian, M.; Julian, B. A.; Mestecky, J.; Novak, J.; Renfrow, M. B. *J. Proteome Res.* **2012**, *11*, 692-702.
- (95) Trinidad, J. C.; Schoepfer, R.; Burlingame, A. L.; Medzihradszky, K. F. *Mol. Cell. Proteomics.* **2013**, *12*, 3474-3488.

- (96) Alley, W. R.; Mechref, Y.; Novotny, M. V. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161-170.
- (97) Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 5023-5032.
- (98) Williams, J. P.; Pringle, S.; Richardson, K.; Gethings, L.; Vissers, J. P. C.; De Cecco, M.; Houel, S.; Chakraborty, A. B.; Yu, Y. Q.; Chen, W. B.; Brown, J. M. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 2383-2390.
- (99) Adamson, J. T.; Hakansson, K. *J. Proteome Res.* **2006**, *5*, 493-501.
- (100) Hakansson, K.; Chalmers, M. J.; Quinn, J. P.; McFarland, M. A.; Hendrickson, C. L.; Marshall, A. G. *Anal. Chem.* **2003**, *75*, 3256-3262.
- (101) Madsen, J. A.; Boutz, D. R.; Brodbelt, J. S. *J. Proteome Res.* **2010**, *9*, 4205-4214.
- (102) Madsen, J. A.; Ko, B. J.; Xu, H.; Iwashiki, J. A.; Robotham, S. A.; Shaw, J. B.; Feldman, M. F.; Brodbelt, J. S. *Anal. Chem.* **2013**, *85*, 9253-9261.
- (103) Zhang, L.; Reilly, J. P. *J. Proteome Res.* **2009**, *8*, 734-742.
- (104) Zhong, Y. Y.; Hyung, S. J.; Ruotolo, B. T. *Expert Rev. Proteomics.* **2012**, *9*, 47-58.
- (105) Damen, C. W. N.; Chen, W. B.; Chakraborty, A. B.; van Oosterhout, M.; Mazzeo, J. R.; Gebler, J. C.; Schellens, J. H. M.; Rosing, H.; Beijnen, J. H. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2021-2033.
- (106) Creese, A. J.; Cooper, H. J. *Anal. Chem.* **2012**, *84*, 2597-2601.
- (107) Woodin, C. L.; Maxon, M.; Desaire, H. *Analyst.* **2013**, *138*, 2793-2803.
- (108) Li, F.; Glinskii, O. V.; Glinsky, V. V. *Proteomics* **2013**, *13*, 341-54.
- (109) Wu, Y.; Mechref, Y.; Klouckova, I.; Mayampurath, A.; Novotny, M. V.; Tang, H. X. *Rapid Commun. Mass Spectrom.* **2010**, *24*, 965-972.
- (110) Wu, S. W.; Liang, S. Y.; Pu, T. H.; Chang, F. Y.; Khoo, K. H. *Journal of Proteomics.* **2013**, *84*, 1-16.
- (111) Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N. *J. Proteome Res.* **2013**, *12*, 3652-3666.

- (112) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (113) Strum, J. S.; Nwosu, C. C.; Hua, S.; Kronewitter, S. R.; Seipert, R. R.; Bachelor, R. J.; An, H. J.; Lebrilla, C. B. *Anal. Chem.* **2013**, *85*, 5666-5675.
- (114) Cooper, C. A.; Gasteiger, E.; Packer, N. H. *Proteomics.* **2001**, *1*, 340-349.
- (115) Darula, Z.; Chalkley, R. J.; Baker, P.; Burlingame, A. L.; Medzihradszky, K. F. *Eur. J. Mass Spectrom.* **2010**, *16*, 421-428.
- (116) Darula, Z.; Chalkley, R. J.; Lynn, A.; Baker, P. R.; Medzihradszky, K. F. *Amino Acids.* **2011**, *41*, 321-328.
- (117) Zhu, Z.; Su, X.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 8403-8411.
- (118) Mayampurath, A.; Yu, C. Y.; Song, E. W.; Balan, J.; Mechref, Y.; Tang, H. X. *Anal. Chem.* **2014**, *86*, 453-463.
- (119) North, S. J.; Hitchen, P. G.; Haslam, S. M.; Dell, A. *Curr. Opin. Struct. Biol.* **2009**, *19*, 498-506.
- (120) Burlingame, A. L. *Curr. Opin. Biotechnol.* **1996**, *7*, 4-10.
- (121) Myers, S. A.; Daou, S.; Affar, E. B.; Burlingame, A. *Proteomics.* **2013**, *13*, 982-991.
- (122) Mayampurath, A. M.; Wu, Y.; Segu, Z. M.; Mechref, Y.; Tang, H. X. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 2007-2019.
- (123) Schedin-Weiss, S.; Winblad, B.; Tjernberg, L. O. *Febs J.* **2014**, *281*, 46-62.
- (124) Sun, S. S.; Wang, Q. Z.; Zhao, F.; Chen, W. T.; Li, Z. *PLoS One.* **2011**, *6*, 1-9.
- (125) Go, E. P.; Irungu, J.; Zhang, Y.; Dalpathado, D. S.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. *J. Proteome Res.* **2008**, *7*, 1660-1674.
- (126) Wuhler, M.; Porcelijn, L.; Kapur, R.; Koeleman, C. A. M.; Deelder, A. M.; de Haas, M.; Vidarsson, G. *J. Proteome Res.* **2009**, *8*, 450-456.

CHAPTER II

GlycoPep Detector: A tool for assigning N-linked glycopeptides based on their ETD-MS/MS spectra

This work has been published by the journal Analytical Chemistry, with reprint permission from the journal.

Electron transfer dissociation (ETD) is commonly used in fragmenting N-linked glycopeptides in their mass spectral analyses to complement collision induced dissociation (CID) experiments. The glycan remains intact through ETD, while the peptide backbone is cleaved, providing the sequence of amino acids for a glycopeptide. Nonetheless, data analysis is a major bottleneck to high throughput glycopeptide identification based on ETD data, due to the complexity and diversity of ETD mass spectra compared to CID counterparts. GlycoPep Detector (GPD) is a web-based tool to address this challenge. It filters out noise peaks that interfere with glycopeptide sequencing, correlates input glycopeptide compositions with the ETD spectra, and assigns a score for each candidate. By considering multiple ion series (c- , z- and y-ions) and scoring them separately, the software gives more weighting to the ion series that matches peaks of high intensity in the spectra. This feature enables the correct glycopeptide to receive a high score while keeping scores of incorrect compositions low. GPD has been utilized to interpret data collected on six model glycoproteins (RNase B, avidin, fetuin, asialofetuin, transferrin and AGP) as well as a clade C HIV envelope glycoprotein, C.97ZA012 gp140ΔCFI. In every assignment made by GPD, the correct glycopeptide composition earns a score that is at least two-fold higher than other incorrect glycopeptide candidates (decoys). The software can be accessed at <http://glycopro.chem.ku.edu/ZZKHome.php>.

2.1 Introduction

Protein glycosylation, one of the most prevalent and fundamental post-translational modifications, plays a regulatory role in numerous cellular activities such as fine-tuning of protein structure and function,¹⁻² antigen-antibody recognition,³⁻⁴ and cell signaling.⁵ Moreover, distinct glycosylation profiles of glycoproteins at normal and disease states are potential targets in biomarker discovery.⁶⁻⁷ Therefore, it is essential to determine glycan structures associated with the modified proteins in order to fully understand their biological significance.

Glycan analysis, while useful for determining glycan structure and heterogeneity, lacks the ability to obtain information regarding the glycan attachment site because the carbohydrates need to be detached from the proteins prior to analysis.⁶⁻⁹ Conversely, in glycopeptide studies by mass spectrometry (MS), individual glycans are retained on peptide chains after proteolysis so that site-specific glycosylation can be studied using tandem MS techniques (MS/MS).¹⁰⁻¹¹

Collisional activation methods including collision-induced dissociation (CID) and infrared multiphoton dissociation (IRMPD) have been extensively applied in fragmenting glycopeptides and characterizing their compositions.¹²⁻¹⁴ Under these conditions, glycosidic bonds are preferentially cleaved to provide compositional information of the glycan moiety while the peptide part remains intact. Several bioinformatics tools have been designed to automate the data-processing procedure of glycopeptide CID-MS/MS spectra.¹⁵⁻²¹ For example, Woodin *et al.* developed a freely-accessible program that utilizes observed dissociation patterns of *N*-linked glycopeptides in CID to score input glycopeptide compositions.²¹ Other software can also characterize the fragmentation of glycans on glycopeptides if the peptide portion is already known.²²⁻²³ However, peptide backbone cleavages of a glycopeptide are severely limited in CID, making it difficult to determine the glycopeptide sequence. In other words, one cannot differentiate two glycopeptide compositions that have the same glycan portion and isobaric (yet

different) peptide sequences based on CID spectra, nor is it possible to locate the *N*-glycosylation site if more than one potential site exists in a single sequence.^{21, 24} These limitations are alleviated by employing alternative fragmentation tools, primarily electron capture dissociation (ECD) and electron transfer dissociation (ETD), to generate c- and z-type peptide backbone fragments while leaving the glycan portion unbroken.^{13, 25-27} The resulting MS/MS spectra are distinct from CID data with no oxonium ions (*m/z* 366, 657, *etc.*) present as indicators of glycopeptide species, and few if any ions that correspond to neutral losses of monosaccharides. While ETD spectra are complementary to CID data, the current biggest shortcoming of utilizing ETD spectra in analyzing glycopeptides is that the spectra need to be interpreted manually.

Herein we introduce GlycoPep Detector (GPD), a freely available software that has been uniquely designed to analyze ETD-MS/MS spectra for *N*-linked glycopeptide identification. Using the program, typical non-informative peaks from electron transfer reactions (charge-reduced species, *etc.*) are removed while other signals are amplified in intensity to reduce random matches.²⁸⁻²⁹ Different ion series (c-, z- and y-ions) from an input glycopeptide composition are searched against the spectrum independently and their weightings for the final score are determined by the sum of intensity of the matched peaks for respective ion types. This unique weighting feature was implemented to account for glycopeptides' distinctive ETD fragmentation patterns, in that typically either c- or z-ions are detected, but the dominant ion series is not readily predictable in advance. Doubly charged fragment ions (c^{2+} -, z^{2+} -ions) are also taken into account for precursor ions at 4+ charge state and above, therefore maximizing the number of informative ions that are used for assignment. The novel algorithm proved to be powerful in differentiating correct glycopeptide candidates from decoys with high specificity. This web-based tool will greatly facilitate data analysis workflows in the glycoproteomics field. It can be accessed at <http://glycopro.chem.ku.edu/ZZKHome.php>.

2.2 Experimental Procedures

2.2.1 Materials and Reagents

Bovine ribonuclease B (RNase B), avidin, fetuin, asialofetuin, human serum glycoproteins (transferrin, AGP) and Sepharose CL-4B were purchased from Sigma Aldrich (St. Louis, MO). Sequencing grade trypsin was acquired from Promega (Madison, WI). Chemical reagents were of analytical purity or better.

2.2.2 Glycoprotein Digestion

Glycoprotein samples (72-300 µg) were dissolved in 100 mM Tris-HCl, pH 8.0 buffer containing 6 M urea for denaturation. Disulfide bonds of the proteins were reduced with 5 mM tris(2-carboxyethyl)phosphine (TCEP) and alkylated with 10 mM iodoacetamide (IAM) for 1 h at room temperature in the dark. The alkylation reaction was quenched by adding 10 mM dithiothreitol (DTT). The Tris-HCl buffer was added to dilute the urea concentration to 1 M, followed by the addition of trypsin at a 1:30 enzyme to protein ratio (w/w). Samples were incubated at 37 °C for 18 h, and the protease digestion was terminated by adding 1 µL acetic acid. The digested solutions were directly analyzed by LC-MS except for the RNase B and avidin samples, which were subjected to hydrophilic enrichment of glycopeptides using Sepharose beads prior to MS analysis.³⁰⁻³¹

2.2.3 Glycopeptide Enrichment

The digested sample (RNase B, avidin) was added to 25 µL Sepharose beads mixed with 1 mL of washing solution (5:1:1 v/v of butanol/ethanol/water). The mixture was shaken gently for 45 min and then centrifuged to discard the supernatant. The wash step was repeated twice to wash

the peptides off the beads. The glycopeptides bound to the beads were then extracted by the addition of 1mL elution solution (1:1 ethanol/water). The extraction was repeated two more times and the supernatant was collected and combined. The combined solution was dried in a Labconco Centrivap cold trap (Kansas City, MO) and reconstituted in 100 μ L 1:1 water/methanol with 0.5% acetic acid. The reconstituted RNase B sample was analyzed by a direct injection experiment, while the avidin sample was analyzed by LC-MS.

2.2.4 Direct Injection Mass Spectrometry

The purified RNase B sample with a concentration of 10 μ M was directly injected into a LTQ Velos linear ion trap (ESI-LIT) mass spectrometer (ThermoScientific, San Jose, CA) at a flow rate of 3 μ L/min in the positive ion mode. The spray voltage was optimized at 3.0 kV and the carrier gas, N₂, was set to 10 psi. The capillary temperature was set as 200 °C. For a selected precursor ion, both CID- and ETD-MS/MS experiments were performed in which the precursor ion was isolated in a 2 Da isolation window. For CID, the activation time was set as 30 ms and the activation energy was 30%. For ETD, the maximum injection time of fluoranthene anions was 150 ms and the reaction time was set as 100 ms with supplemental activation turned on.³² The MS/MS spectra were recorded by averaging 30 scans with each scan containing 10 microscans.

2.2.5 Liquid Chromatography/Mass Spectrometry

LC-MS analysis was conducted using a C₁₈ column (300 μ m i.d. \times 5 cm, 100 Å pore size, Micro-Tech, Vista, CA) coupled to a LTQ Velos linear ion trap mass spectrometer via an ACQUITY UPLC system (Waters, Milford, MA). Five microliters of a sample was injected for one run. Different gradients were used for optimized separations of different samples. For fetuin, asialofetuin, transferrin and AGP, the column was flushed with 98% eluent A (99.9% H₂O with

0.1% formic acid) and 2% eluent B(99.9% CH₃CN with 0.1% formic acid) for 5 min at a flow rate of 7 μ L/min, followed by a linear increase of eluent B to 40% in 50 min. For avidin, the gradient started with 2% B for 10 min and was ramped to 40% B in 35 min. After the separation gradient, the column was washed with 90% eluent B for 10 min and was subsequently re-equilibrated with 2% eluent B for another 10 min.

The mass spectrometer was operated using the following conditions: the spray voltage of the ESI source was 3.0 kV and the capillary temperature was set as 200 °C. Each sample was analyzed in two separate runs that were set for CID and ETD experiments, respectively. The MS/MS spectra were collected in a data-dependent mode. Five most intense ions in the full scan (m/z of 500-2000) were sequentially selected for CID or ETD with a 3 min dynamic exclusion window. Normalized collision energy of 30% and activation time of 10 ms were set for each CID scan. For ETD settings, the maximum injection time of fluoranthene anions was 150 ms, and the reaction time was 90 ms with supplemental activation turned on.

2.2.6 Glycopeptide Training Dataset

The ETD-MS/MS spectra of glycopeptides with known peptide sequences and glycan compositions were needed to reveal the fragmentation patterns for algorithm development. MS and MS/MS data were collected from trypsinized glycoproteins (RNase B, avidin, fetuin, asialofetuin, transferrin and AGP) that have been studied extensively.^{27, 31, 33-35} The glycoprotein sequences were *in silico* digested to generate tryptic peptides with up to 2 missed cleavages allowed, and all the cysteine residues were carbamidomethylated. The predicted glycopeptide masses were calculated by summing up the masses of the peptides that contain the *N*-glycosylation sites and the masses of the reported glycan compositions for each glycosylation site. The theoretical m/z values of each glycopeptide composition were then computed, and the full scan

spectrum was searched in 1 min increments for MS¹ peaks within 200 ppm of the m/z values of the predicted glycopeptides. If a match was found, the CID-MS/MS spectrum was scrutinized to confirm the glycopeptide composition that constituted the MS¹ peak. The corresponding ETD-MS/MS spectrum of the confirmed glycopeptide peak was then carefully analyzed to evaluate and verify the glycopeptide fragmentation in ETD.

2.2.7 Glycopeptide Validation Dataset

After the GPD program was completed, an additional test was conducted to validate the performance of the software in interpreting glycopeptide ETD-MS/MS spectra. The ETD data was collected from a clade C HIV envelope glycoprotein, C.97ZA012 gp140 Δ CFI, which had been characterized by MS and CID-MS/MS before.³⁶ This protein has 24 potential *N*-glycosylation sites with over 90% sites occupied by diverse glycan compositions. Consequently, the glycopeptides originating from this protein are heterogeneous in both the peptide sequence and the glycosylation profile. Every ETD spectrum was analyzed manually to verify the glycopeptide composition assigned by the GPD software, and the MS and CID-MS/MS data were utilized to further confirm the assignment.

2.2.8 Glycopeptide Data Input

The ETD-MS/MS spectrum is exported as a peak list file in .CSV format, which is uploaded to GPD. The m/z and charge state of the precursor ion are needed, along with the lower and upper m/z limit of the MS/MS scan. GPD calculates the m/z values of the charge-reduced precursors within the scan range and removes interfering peaks from the precursor ion and charge-reduced precursors. Three additional parameters are manually input: N1, N2 and K. N1 is the number of peaks that are preserved by intensity in every 100 Da interval for the low m/z region

(from the lower scan limit to the precursor m/z), while N_2 is the number of peaks that are preserved in each interval for the high m/z region (above the precursor m/z). K is the amplification factor of the remaining peaks in the low m/z region. These parameters are set for the data pre-processing steps to reduce the noise peaks while amplifying signal peaks, as is described in the discussion section.

The glycopeptide candidate compositions are input manually with each taking a separate line. Three sections need to be entered for one candidate: the peptide portion, the glycan portion (in the form of $[\text{Hex}]_n[\text{HexNAc}]_n[\text{Neu5Ac}]_n[\text{Fuc}]_n$), and the glycosylated asparagine index that indicates the glycosylation site in cases where more than one asparagine is in the sequence. GPD computes and searches for the m/z of c-, z-, and y-ions derived from each input composition through the peak list. If the precursor ion is at 4+ or higher charge state, the c^{2+} - and z^{2+} -ions are also searched. After correlating the input composition to the uploaded data, GPD assigns a final score for each glycopeptide candidate.

2.2.9 False Discovery Rate Determination

To evaluate the false discovery rate of the GPD program, decoy glycopeptide compositions were generated from an in-house database. A decoy protein, *Titin*, which consists of 50,000 amino acid residues, was *in silico* digested to generate peptide sequences that were used as the peptide portions for the decoy glycopeptides.²¹ The glycan portions of the decoys came from a library of around 200 biologically relevant carbohydrate compositions.³⁷ To examine the program's ability to differentiate the right composition from multiple decoys, a relatively large mass error window of 200 ppm was set for selecting decoy glycopeptide candidates. For every spectrum analyzed by GPD, a minimum of 4 decoy candidates were entered along with the correct glycopeptide composition to estimate the false-positive rate of the software.

2.3 Results and Discussion

The ultimate goal of this work is to design and test a web-based analysis tool that assigns ETD spectra of glycopeptides in a highly accurate and automated fashion. To accomplish this objective, a library of ETD data of known glycopeptides was required. This library was used to identify important trends in terms of ETD fragmentation of glycopeptides that could then be exploited in an assignment algorithm. After algorithm development, validation could be accomplished using glycopeptide data from a distinct set of ETD spectra.

2.3.1 Spectral Library Generation

The model glycoproteins selected include RNase B, avidin, fetuin, asialofetuin, transferrin and AGP. These proteins contain diverse glycoforms, and their glycopeptides have been extensively analyzed and previously described. Glycopeptides from these trypsinized glycoproteins were subjected to both CID- and ETD-MS/MS analysis, and their CID data were employed to verify every glycopeptide composition assignment made prior to analyzing their ETD spectra.

2.3.2 Characteristics of Glycopeptide Fragmentation in ETD

It is known that peptide backbone dissociation in ETD is more efficient for precursor ions at higher charge states,³⁸⁻³⁹ and the same trend has been observed for glycopeptides.²⁷ Figure 1 summarizes ETD spectra of a 32-amino acid-long glycopeptide from asialofetuin with 4+, 5+ and 6+ charges (Figure 1A, B and C). No backbone cleavage occurred for the 4+ glycopeptide ion of m/z 1415.6 (Figure 1A), and it was reported previously that ETD has a useful m/z range of less than 1400.²⁷ As the charge increases, glycopeptide backbone fragments appear and become more pronounced (Figure 1B and C), which is consistent with the prior research for peptides.³⁸ Since

electron transfer reaction reduces the overall charge, the precursor ion fragments most efficiently at higher charge states. In our analyses, glycopeptides at 3+ and higher charge state with less than 1400 m/z value generated most informative ETD data that could be employed for compositional assignment, hence we focused on developing a tool to primarily analyze glycopeptide ETD spectra that fell into this category.

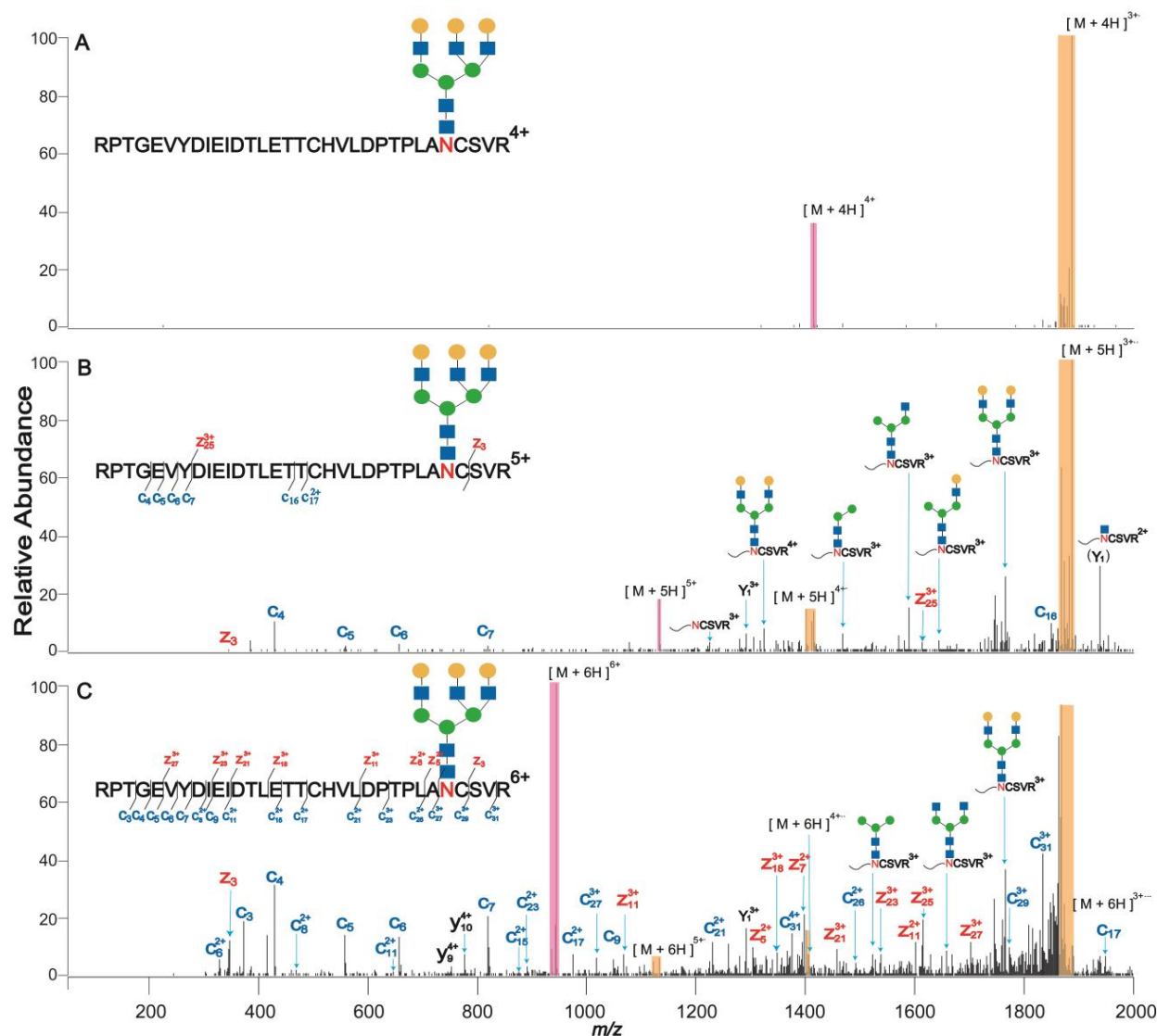


Figure 1. ETD-MS/MS data of a 32-amino acid-long glycopeptide with a tri-antennary complex *N*-glycan of asialofetuin at 4+ (A, m/z 1415.6), 5+ (B, m/z 1132.7) and 6+ (C, m/z 944.1) charge states. Precursor ion peaks are marked with purple bars, while charged reduced species and their neutral losses are marked in yellow bars. Products from glycosidic bond cleavages are also shown. Different types of peptide backbone fragments (c-, z-, y-ions) are labeled in different colors as indicated in the figure. Glycan symbols include,

blue square: *N*-acetylhexosamine, green and yellow circle: hexose, purple diamond: sialic acid. Subsequent figures are illustrated in the same way unless otherwise noted.

After an extensive library of over 90 glycopeptide ETD spectra was acquired, the library was used to assign key features of glycopeptide fragmentation that could be utilized upon algorithm development. Distinct from ETD data of peptides, only one ion series, of either c- or z-type ions, was found to be dominant in the ETD spectra of various glycopeptides. The loss of one of the two ion series for glycopeptides can sometimes be attributed to the additional mass that the glycan adds to the fragment ions. As is demonstrated in Figure 2A, which shows the ETD spectrum of a glycopeptide from transferrin, singly charged c-ions dominate over other ions while essentially no z-ions are observed. This trend is caused by the *N*-glycan modification (+2204.8 Da) that is adjacent to the C-terminus. The glycan shifts all z-ions (except z_1^+) beyond the MS/MS scan range (m/z 50-2000) while c-ions are largely unaffected, because they do not contain the glycosylated asparagine (except c_{12}^+). In contrast, Figure 2B shows this same glycopeptide with a missed tryptic cleavage. The missed cleavage on the C-terminus of CGLVPVLAENYNK led to an extension of 20 amino acid residues and changed the relative location of the glycosylation site, making it closer to the N-terminus instead. As a result, an extensive z-ion series is identified while few c-ions are recorded in Figure 1B. This observation conforms to the fact that for the glycopeptide with the missed cleavage, most of its z-ions do not contain the glycan portion.

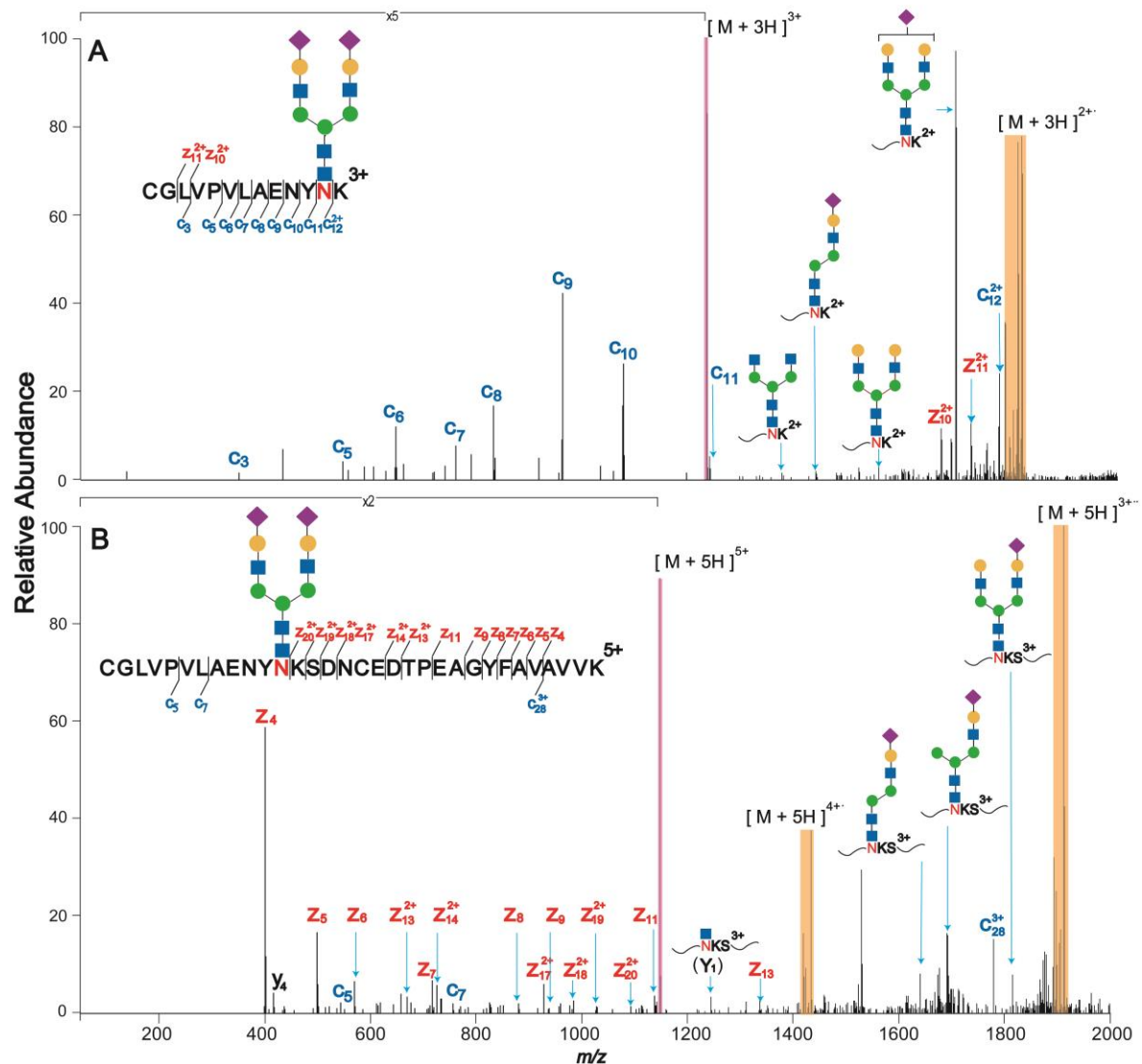


Figure 2. ETD-MS/MS data of (A) a tryptic glycopeptide with a bi-antennary complex N -glycan from transferrin ($3+$, m/z 1227.8) and (B) the same glycopeptide species but with one missed cleavage ($5+$, m/z 1147.7).

The additional mass of the glycan is likely not the only contributing factor to why only one of the two ion series are typically observed for ETD of glycopeptides. Even for glycopeptides with highly similar structures, their ETD fragmentation patterns can differ, and sometimes, only one dominant ion series is present. Figure 3 consists of two spectra collected from homologous high mannose-type glycopeptides that differ by one mannose residue (Man7 vs. Man6). Both c - and z -ions are produced in the Man7 glycopeptide spectrum, though c -ions are in lower intensity than

z-ions (Figure 2A). Nonetheless, in the Man6 spectrum, Figure 2B, only the z-ion series exist while c-ions are completely missing. The discrepancy between these ETD spectra may be due to the different abundance of the two glycoforms, since the normalized intensity of the Man7 glycopeptide was about 5 times higher than that of the Man6 glycopeptide in the full scan spectrum (data not shown). However, regardless of its cause, the unique phenomenon that only one ion series is frequently generated in the ETD spectrum of a glycopeptide needs to be accounted for by an effective algorithm designed for the GPD program.

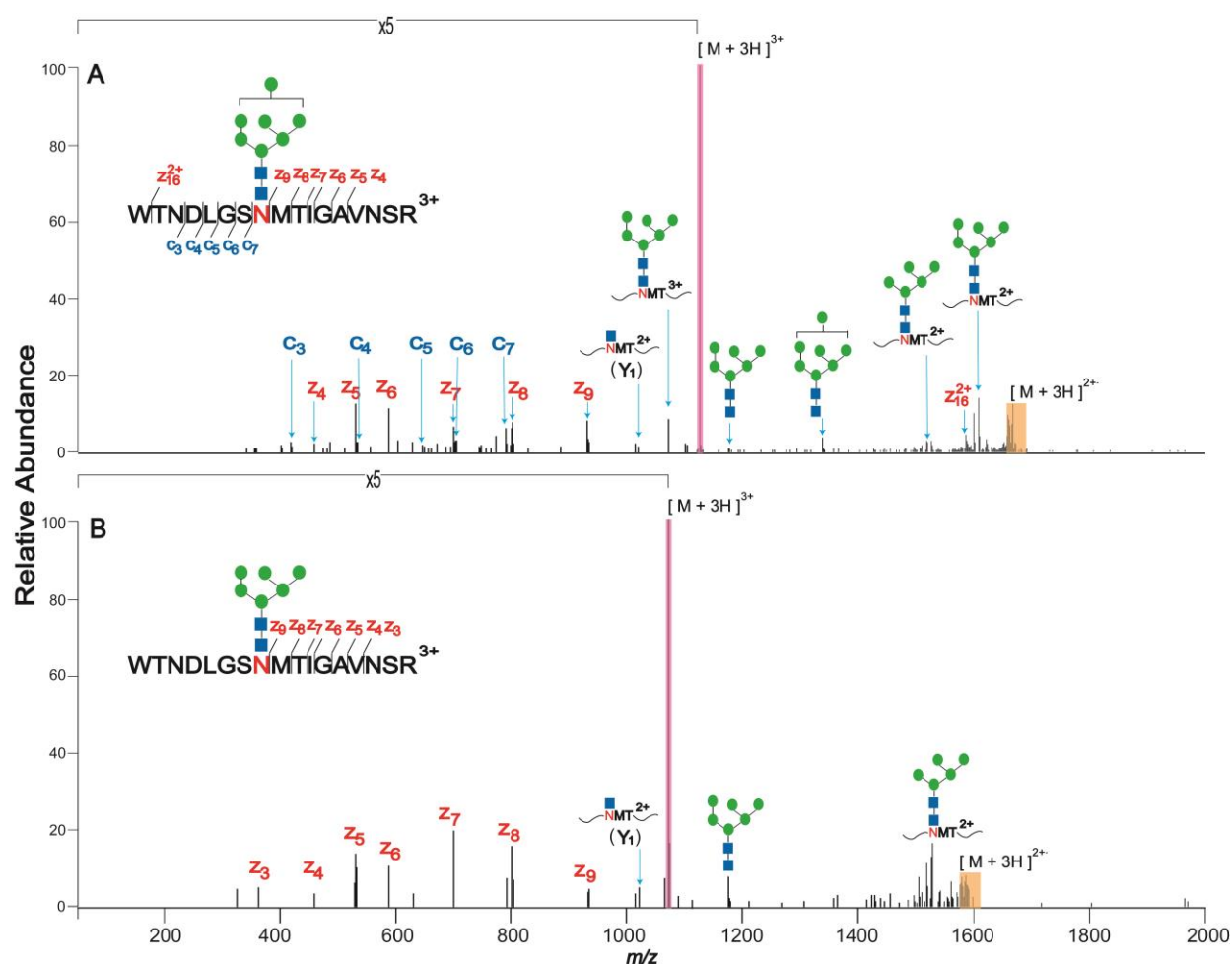


Figure 3. ETD data from two avidin glycopeptides at 3+ charge state with the same peptide sequence and homologous high mannose *N*-glycans attached: Man7 (A, m/z 1126.1) and Man6 (B, m/z 1072.1) tryptic glycopeptides.

We also found that even-electron $z+1$ ions were generated frequently in glycopeptide ETD spectra, and a larger fragment ion mass window was used for z -ions to incorporate $z+1$ ions. In

contrast, odd-electron c-1 ions were rarely observed in our dataset, which is in agreement with previous reports that few c-1 ions exist in the spectra of precursors with 3+ or higher charges.³⁹⁻⁴⁰ Peaks from cleavage on the glycan portion of the glycopeptide ion also appeared in some spectra, but the fragmentation was limited and these ions were not searched for in the algorithm.

2.3.3 Data Pre-processing of the Raw Spectrum

As is exemplified in Figure 1C, numerous peaks that do not reveal the glycopeptide sequence exist in the glycopeptide ETD spectrum. In addition, these peaks of noise are not evenly distributed through the m/z range. In the low m/z region of the spectrum, interfering peaks are in low abundance compared to the signal peaks from glycopeptide backbone fragments. On the contrary, peaks arising from charge-reduced species and glycosidic bond cleavages that interfere with sequencing are dominant in high m/z area. This feature prompted us to make software that processes peaks located in different parts of a spectrum differently in order to eliminate noise peaks while maximizing the number of signal peaks.

As is described above, interfering peaks from side products of electron transfer reactions and glycan dissociations in the raw spectrum need to be filtered out to improve match confidence. The GPD program employs a stepwise processing method to perform spectral filtering before a spectrum is scored, and the procedure is illustrated in Figure 4. Firstly, the precursor ion peak, charge-reduced precursors and their neutral losses in the raw spectrum (Figure 4A) are deleted by an approach similar to that of processing peptide ETD spectra.²⁸ Isotope replicates in the resulting spectrum (Figure 4B) are then eliminated by removing all peaks except the tallest one in each 3 Da bin to generate data shown in Figure 4C. Next, the spectrum is divided into two areas by the precursor m/z : for region 1, from the lower scan limit to the precursor m/z , the top 5 peaks of highest intensity in every 100 Da interval are preserved with other peaks removed; for region 2,

which is above the precursor m/z , only the 3 tallest peaks are reserved in each interval. Finally, the intensity values of the remaining peaks in the first region are amplified by a factor of 5. The completely processed spectrum is present in Figure 4D, which can then be subjected to spectral scoring. Following these steps, signal peaks with good S/N in the low m/z end are preserved and weighted more heavily, even if their intensities are low compared to noise peaks in the high m/z region of the raw spectrum.

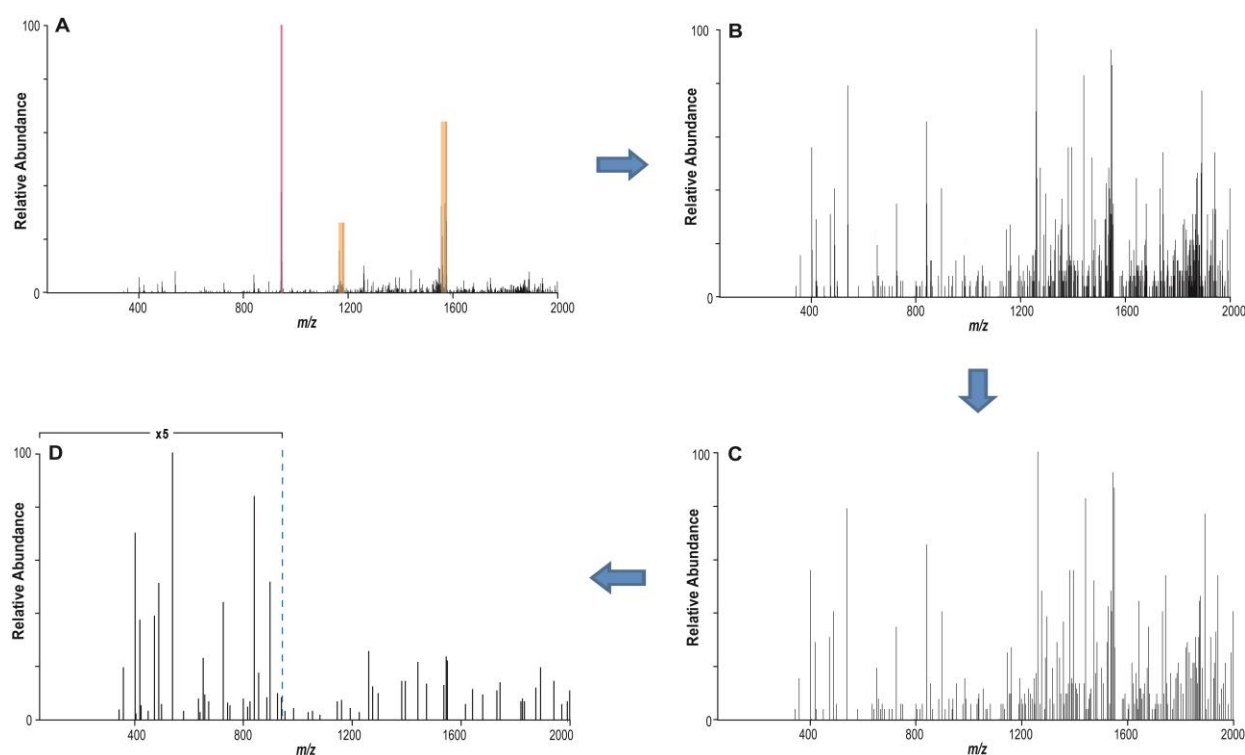


Figure 4. Scheme of the spectral pre-processing method. A raw glycopeptide ETD spectrum is shown in (A). In the first step, the precursor ion, charge-reduced precursors and the neutral losses are removed to generate the data in (B). Redundant isotope peaks were then eliminated and the processed spectrum was illustrated in (C). In the last step, peaks of highest intensity in each 100 Da interval were preserved while other peaks are removed, and the remaining peaks in the low m/z region are further amplified in intensity. The final spectrum after the pre-processing steps are demonstrated in (D). The base peak is re-scaled after each step.

2.3.4 Novel Aspects of GPD Algorithm

Since the *N*-glycan side chain typically causes only one ion series to be well represented in the spectrum, the peptide backbone fragments (c-, z- and y-ions) are independently tracked. The *m/z* of every ion in each ion series is computed based on the input glycopeptide composition, and these *m/z* values are searched against the spectral data. An individual score for each ion series is calculated based on the number of matched ions out of all possible ions using a binomial distribution, as is shown below:

$$\text{Score}(k - \text{ion}) = -10 \times \log[\sum_{k=n}^N \binom{N}{k} p^k (1-p)^{N-k}]$$

In the equation, *N* and *n* denote the total number of possible *k*-ions and the number of *k*-ions matched to the spectrum, respectively, and *p* is the probability of randomly matching one ion to the spectrum. Subsequently, each ion series' score is weighted according to the sum of intensities of spectral peaks assigned to that specific ion series with weighting factors normalized to 1. The final score is thus computed by the following equation:

$$\text{Final Score} = \sum_{k=c, z, y} \left[\frac{\sum \text{Int.}(k\text{-ions})}{\sum \text{Int.}(\text{all ions})} \times \text{Score}(k - \text{ion}) \right]$$

For precursor ions with 4+ charges and above, c²⁺- and z²⁺-ion series are also included in the equation. Using this unique scoring algorithm, the absence of c- or z-ions does not impact the final score of the correct glycopeptide composition, because the contribution from absent ion series with low individual scores would be minimal. As for decoy candidates, the match between a theoretical and experimental spectral pair occurs randomly and is evenly distributed among different ion series. The algorithm utilizes this feature to keep the final score of the decoy composition low by considering multiple types of ions and abrogating abnormal scores of individual ion series.

2.3.5 GPD Scoring of ETD Spectra from Model Glycoproteins

Glycopeptide ETD-MS/MS data from model glycoproteins, our glycopeptide training dataset, was submitted to GPD software for spectral processing and scoring. For each spectrum, the correct glycopeptide composition, as well as at least four decoy glycopeptides, were entered and scored.

The scoring of the ETD spectrum from a sialylated complex type glycopeptide is present in Figure 5. It should be noted that the spectrum has gone through data pre-processing as described above. The processed spectrum is scored against the correct composition (Figure 5A) along with a decoy that has the same glycan portion ([Hex]5[HexNAc]4[Neu5Ac]2) and an isobaric peptide sequence (Figure 5B). Peaks that are matched to different ion series are in different colors as indicated in the figure. The parameters to calculate the final score of each candidate are listed in Table 1. For the correct assignment, signal peaks in the low m/z region that are significantly higher than other peaks after spectral processing are mostly assigned as c- and z-ions, which get superior ion series scores of 43.1 and 56.2 because they fit the processed spectrum with high specificity. In contrast, other ion series do not have such high sequence coverage, and their scores are lower, especially for z^{2+} -ions where only z_7^{2+} is matched to receive a low score of 2.4. Nonetheless, the sum of the weightings for c- and z-ions is 80% as compared to that of z^{2+} -ion series that only weights for 0.5% to give a high final score of 44.2. For the decoy assignment, each ion series only has up to two matches to the spectral peaks in a random way so that all of their individual scores are less than 4.5, and the final score of the decoy is 4.1, which is ten times lower than that of the real composition.

Table 1. GPD scoring parameters of the correct glycopeptide composition and a decoy candidate that are shown in Figure 5.

Correct composition: QQQHLFGS N VTDCSGNFCLFR+[Hex]5[HexNAc]4[Neu5Ac]2					
Ion series	c-ion	z-ion	y-ion	c ²⁺ -ion	z ²⁺ -ion
$\frac{\text{\# of matched ions}}{\text{\# of possible ions}}$	5/8	8/12	4/12	4/14	1/13
Ion-series Score	43.1	56.2	22.7	20.1	2.4
Weightings	36.9%	43.1%	7.1%	12.4%	0.5%
Final Score	44.2				
Decoy composition: NEIASHE N VTLPTTPLDPVLAAGR+[Hex]5[HexNAc]4[Neu5Ac]2					
Ion series	c-ion	z-ion	y-ion	c ²⁺ -ion	z ²⁺ -ion
$\frac{\text{\# of matched ions}}{\text{\# of possible ions}}$	1/7	2/13	1/16	1/11	0/10
Ion-series Score	4.3	4.5	1.9	2.9	0
Weightings	9.4%	74.4%	9.4%	6.8%	0%
Final Score	4.1				

A list of GPD final scores is provided in Table S1 where 30 distinct glycopeptides featuring over 90 ETD spectra collected from model glycoproteins were analyzed. The correct composition received the highest score for every tested spectrum, and the score was at least two times higher than that of the best-scored decoy composition.

To evaluate the usefulness of tracking different types of ions independently in the algorithm, ion series' scores of 19 glycopeptides at 4+ and higher charge states were averaged as well as those of the decoy compositions that received the highest final scores among the decoys.

For this subset of data, all five ion series, including the c^{2+} - and z^{2+} -ions, were considered. These data appear in Figure 6A. Error bars of one standard deviation were also included. The ion series' score of the correct assignment is significantly higher than that of the decoy for all five ion series calculated by the program, and the c- and z-ion scores of the correct candidate contrast most to those of the decoy (c-ion score of 35.0 vs. 2.1; z-ion score of 25.8 vs. 3.2). The large standard deviations of the ion series' scores for the correct assignment support our initial findings that individual glycopeptides may fragment by producing *either* numerous c ions *or* numerous z ions; but frequently not both series are well represented. In the algorithm, the contributions from different ion series to the final score are determined in part by the intensity of peaks matched to specific ion types; so for the correct assignment, ion series receiving high individual scores should be given large ratios automatically because the matched peaks would have high intensity. The feature is confirmed by the statistics in Figure 6B where the weightings of each ion series for the same correct and decoy compositions as in Figure 6A were averaged. The ratios of high-scoring c- and z-ion series are predominant (40.2% and 38.7%, respectively) over the other ion series (the most of which is less than 8%) for the correct candidates. In contrast, the weightings of different ion types of the decoys are similar (13.5%-26.5%) due to an even probability of random matches across the different ion series.

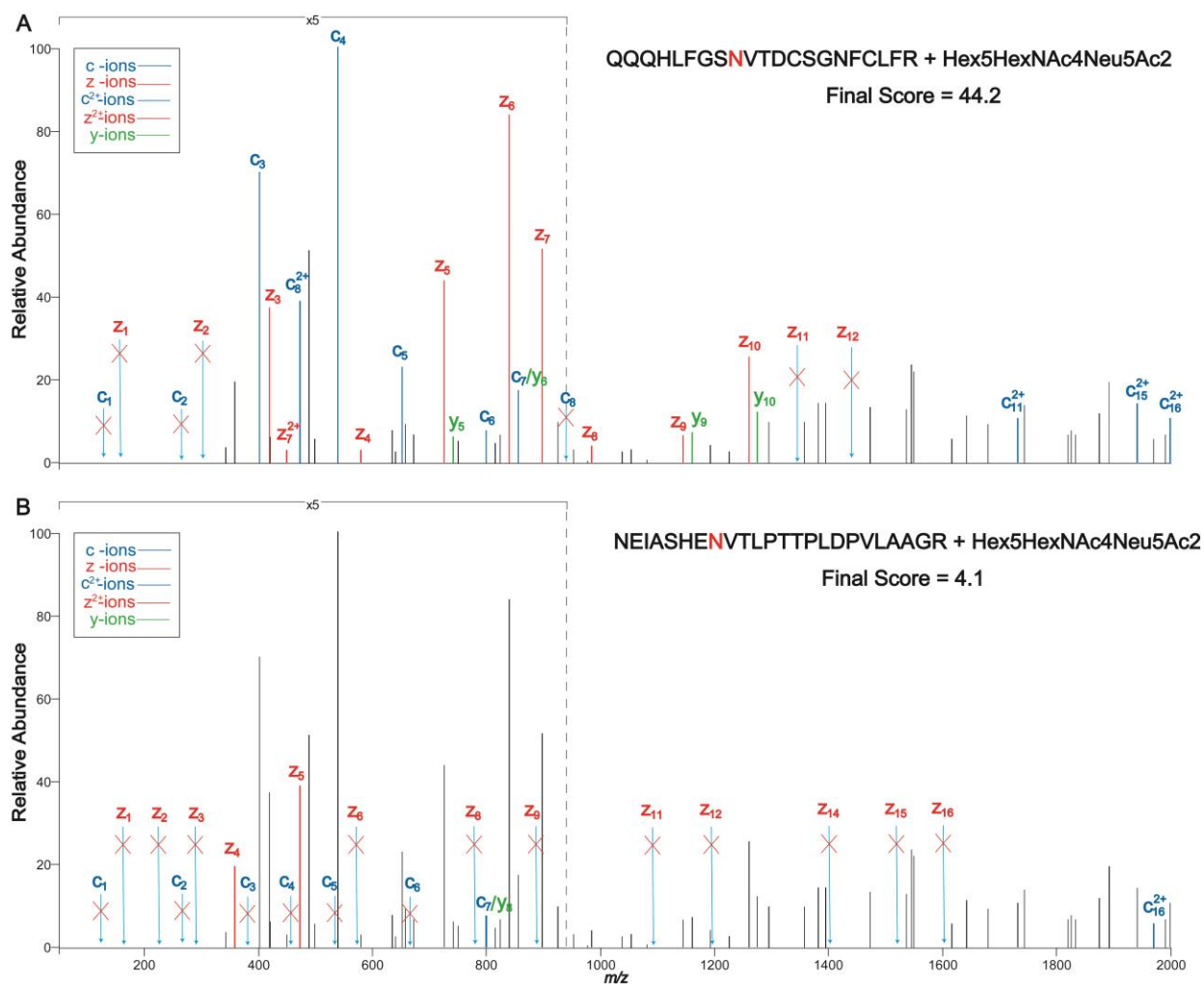


Figure 5. The correct glycopeptide composition, QQQHLFGSNVTDCSGNFCLFR+[Hex]5[HexNAc]4[Neu5Ac]2 (A), was scored against its ETD spectra after spectral cleaning, along with a decoy composition having the same glycan portion but a different peptide sequence of NEIASHENVTLPTTLPDPLAAGR (B). The monoisotopic neutral masses of these two candidates are 4718.8892 and 4719.0802 respectively. Spectral peaks that are matched to different ion series are in different colors as indicated in the figure, while arrows with X's denote that the putative ions are not found in the spectra. The glycosylated asparagines are labeled as red N's for the compositions shown in the figure. The correct composition received a final score of 44.2 (A) while the decoy composition had a final score of 4.1 (B).

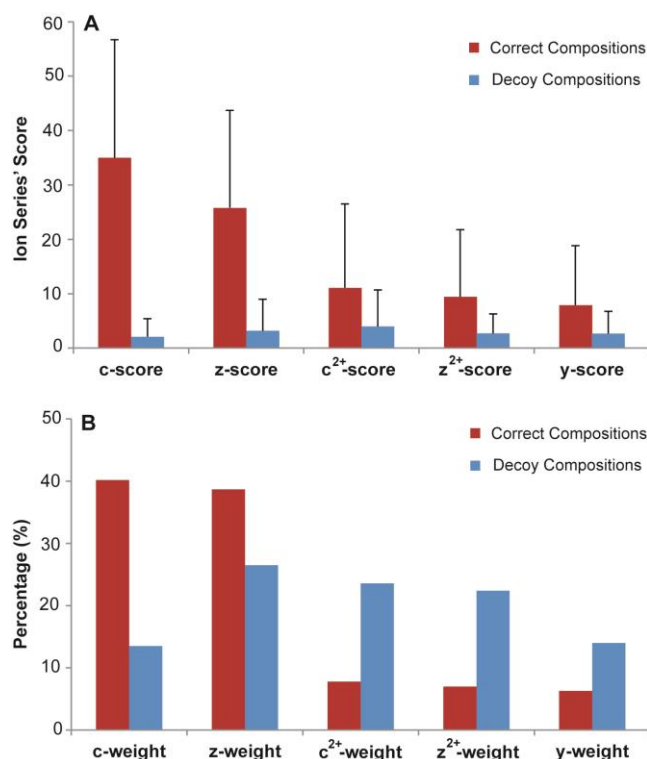


Figure 6. Bar graphs of (A) mean values of ion series' scores of the correct compositions (n=19, shown in red bars), as well as decoys that received the highest final scores (blue bars), with error bars indicating one standard deviation; (B) average weighting of each ion series' score in determining final scores of the glycopeptide compositions analyzed in (A). For glycopeptides carrying more than 3+ charges, five ion types (c-, z-, y-, c²⁺- and z²⁺-ions) are included for scoring.

2.3.6 Extended Test of GPD on ETD Spectra from the HIV Env Glycoprotein, C.97ZA012 gp140ΔCFI

After the initial test of the program on ETD spectra of model glycoproteins, GPD was further employed to score ETD data from tryptic glycopeptides of a clade C HIV envelope protein C.97ZA012 gp140ΔCFI, the glycopeptide validation dataset. We tested more than 120 ETD-MS/MS spectra generated from 45 HIV envelope glycopeptides with 4-7 decoy compositions for each assignment. In addition, to examine if the program is able to differentiate glycopeptide isomers, the sequence of each tested glycopeptide was reversed while the glycan portion is unchanged to generate an isomeric glycopeptide for every test. The results, as illustrated

in Figure 7, indicates that using the GPD program, the correct glycopeptide composition can be easily differentiated from the decoys, including the isomer, because the true candidate received a final score that is at least two fold higher than the decoys in every test. The results of all the tested compositions for the validation dataset are summarized in detail in Table S2.

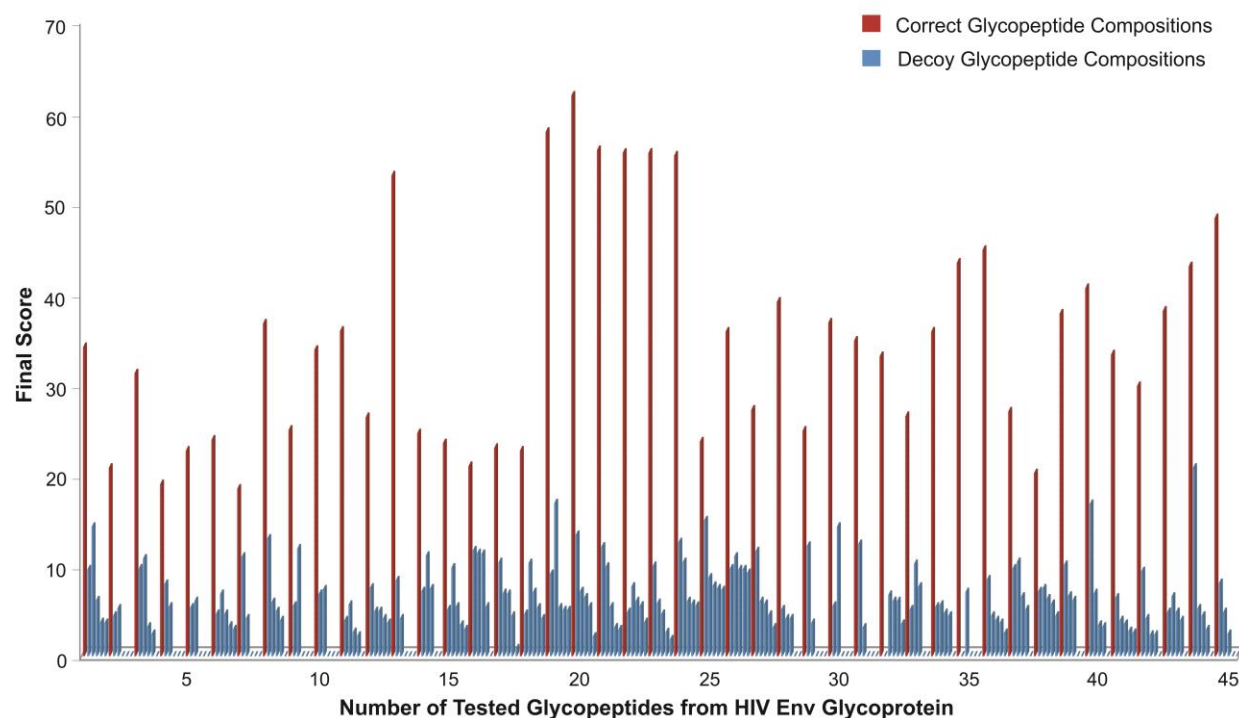


Figure 7. A summary of the final scores of 45 distinct glycopeptides from the HIV Env Glycoprotein (shown in red bars). For each assignment, one glycopeptide isomer with the same glycan portion as the correct composition but reverse in sequence, along with 4 decoy compositions with the highest final scores are also included (as 5 blue bars right next to each corresponding red bar). Note that some decoys receive a final score of 0 and those bars are displayed as blue spots in the figure.

2.4 Conclusion

ETD has become increasingly utilized in proteomics research; and compared to CID, ETD adds an orthogonal dimension in probing glycopeptide structures. Glycopeptide ETD spectra are drastically different from their CID data, because the peptide portion is cleaved in ETD as opposed to the glycan part in CID. Moreover, we also found that ETD data of glycopeptides differ much from those of peptides in that one type of ion series dominates over other ion series. In many cases, this difference is due to the presence of the *N*-glycan modification. This interesting property prompted us to develop a novel algorithm for GlycoPep Detector that scores different ion series independently and weights them by the matched peaks' intensity when a glycopeptide is scored against the spectrum. In addition, the GPD program has also combined a spectral pre-processing method designed for cleaning the raw spectrum prior to scoring in order to extract informative spectral features and eliminate noise peaks.

The web-based analysis tool was highly discriminative towards correct compositions against decoy candidates in tests of glycopeptide ETD-MS/MS spectra collected from tryptic digests of RNase B, avidin, fetuin, asialofetuin, transferrin and AGP and in the extended test for a highly complex sample, an HIV envelope glycoprotein gp140. In sum, 75 unique glycopeptide compositions were correctly assigned by GPD from over 200 ETD spectra, and the correct assignments always received scores at least two times higher than their decoy compositions. The GPD tool is freely available online at <http://glycopro.chem.ku.edu/ZZKHome.php>.

References

- (1) Dwek, R. A. *Science*. **1995**, 269, 1234-1235.
- (2) Kajihara, Y.; Tanabe, Y.; Sasaoka, S.; Okamoto, R. *Chem.-Eur. J.* **2012**, 18, 5944-5953.
- (3) Crouch, E.; Nikolaidis, N.; McCormack, F. X.; McDonald, B.; Allen, K.; Rynkiewicz, M. J.; Cafarella, T. M.; White, M.; Lewnard, K.; Leymarie, N.; Zaia, J.; Seaton, B. A.; Hartshorn, K. L. *J. Biol. Chem.* **2011**, 286, 40681-40692.
- (4) Doores, K. J.; Bonomelli, C.; Harvey, D. J.; Vasiljevic, S.; Dwek, R. A.; Burton, D. R.; Crispin, M.; Scanlan, C. N. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, 107, 13800-13805.
- (5) Furukawa, K.; Ohkawa, Y.; Yamauchi, Y.; Hamamura, K.; Ohmi, Y. *J. Biochem.* **2012**, 151, 573-578.
- (6) Barrabes, S.; Pages-Pons, L.; Radcliffe, C. M.; Tabares, G.; Fort, E.; Royle, L.; Harvey, D. J.; Moenner, M.; Dwek, R. A.; Rudd, P. M.; De Llorens, R.; Peracaula, R. *Glycobiology*. **2007**, 17, 388-400.
- (7) Alley, W. R.; Madera, M.; Mechref, Y.; Novotny, M. V. *Anal. Chem.* **2010**, 82, 5095-5106.
- (8) Hua, S.; An, H. J.; Ozcan, S.; Ro, G. S.; Soares, S.; DeVere-White, R.; Lebrilla, C. B. *Analyst*. **2011**, 136, 3663-3671.
- (9) Leymarie, N.; Zaia, J. *Anal. Chem.* **2012**, 84, 3040-3048.
- (10) Wuhrer, M.; Catalina, M. I.; Deelder, A. M.; Hokke, C. H. *J. Chromatogr. B*. **2007**, 849, 115-128.
- (11) Pan, S.; Chen, R.; Aebersold, R.; Brentnall, T. A. *Mol. Cell. Proteomics*. **2011**, 10, 1-14.
- (12) Huddleston, M. J.; Bean, M. F.; Carr, S. A. *Anal. Chem.* **1993**, 65, 877-884.
- (13) Hakansson, K.; Cooper, H. J.; Emmett, M. R.; Costello, C. E.; Marshall, A. G.; Nilsson, C. L. *Anal. Chem.* **2001**, 73, 4530-4536.
- (14) Seipert, R. R.; Dodds, E. D.; Clowers, B. H.; Beecroft, S. M.; German, J. B.; Lebrilla, C. B. *Anal. Chem.* **2008**, 80, 3684-3692.

- (15) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. *J. Proteome Res.* **2007**, *6*, 3995-4005.
- (16) Ren, J. M.; Rejtar, T.; Li, L. Y.; Karger, B. L. *J. Proteome Res.* **2007**, *6*, 3162-3173.
- (17) Joenvaara, S.; Ritamo, I.; Peltoniemi, H.; Renkonen, R. *Glycobiology.* **2008**, *18*, 339-349.
- (18) Ozohanics, O.; Krenyacz, J.; Ludanyi, K.; Pollreisz, F.; Vekey, K.; Drahos, L. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3245-3254.
- (19) Peltoniemi, H.; Joenvaara, S.; Renkonen, R. *Glycobiology.* **2009**, *19*, 707-714.
- (20) Pompach, P.; Chandler, K. B.; Lan, R.; Edwards, N.; Goldman, R. *J. Proteome Res.* **2012**, 1728-1740.
- (21) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (22) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008**, *7*, 1650-1659.
- (23) Maass, K.; Ranzingei, R.; Geyer, H.; von der Lieth, C. W.; Geyer, R. *PROTEOMICS.* **2007**, *7*, 4435-4444.
- (24) Go, E. P.; Hewawasam, G.; Liao, H. X.; Chen, H. Y.; Ping, L. H.; Anderson, J. A.; Hua, D. C.; Haynes, B. F.; Desaire, H. *J. Virol.* **2011**, *85*, 8270-8284.
- (25) Hogan, J. M.; Pitteri, S. J.; Chrisman, P. A.; McLuckey, S. A. *J. Proteome Res.* **2005**, *4*, 628-632.
- (26) Catalina, M. I.; Koeleman, C. A. M.; Deelder, A. M.; Wuhler, M. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 1053-1061.
- (27) Alley, W. R.; Mechref, Y.; Novotny, M. V. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161-170.
- (28) Good, D. M.; Wenger, C. D.; McAlister, G. C.; Bai, D. L.; Hunt, D. F.; Coon, J. J. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1435-1440.
- (29) Renard, B. Y.; Kirchner, M.; Monigatti, F.; Ivanov, A. R.; Rappsilber, J.; Winter, D.; Steen, J. A. J.; Hamprecht, F. A.; Steen, H. *PROTEOMICS.* **2009**, *9*, 4978-4984.

- (30) Wada, Y.; Tajiri, M.; Yoshida, S. *Anal. Chem.* **2004**, *76*, 6560-6565.
- (31) Rebecchi, K. R.; Wenke, J. L.; Go, E. P.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1048-1059.
- (32) Swaney, D. L.; McAlister, G. C.; Wirtala, M.; Schwartz, J. C.; Syka, J. E. P.; Coon, J. J. *Anal. Chem.* **2007**, *79*, 477-485.
- (33) Zhang, Y.; Go, E. P.; Desaire, H. *Anal. Chem.* **2008**, *80*, 3144-3158.
- (34) Liu, X.; McNally, D. J.; Nothaft, H.; Szymanski, C. M.; Brisson, J. R.; Li, J. J. *Anal. Chem.* **2006**, *78*, 6081-6087.
- (35) Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; Kawasaki, N.; Khoo, K. H.; Kim, S.; Kondo, A.; Lattova, E.; Mechref, Y.; Miyoshi, E.; Nakamura, K.; Narimatsu, H.; Novotny, M. V.; Packer, N. H.; Perreault, H.; Peter-Katalinic, J.; Pohlentz, G.; Reinhold, V. N.; Rudd, P. M.; Suzuki, A.; Taniguchi, N. *Glycobiology*. **2007**, *17*, 411-422.
- (36) Go, E. P.; Chang, Q.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. *J. Proteome Res.* **2009**, *8*, 4231-4242.
- (37) Go, E. P.; Rebecchi, K. R.; Dalpathado, D. S.; Bandu, M. L.; Zhang, Y.; Desaire, H. *Anal. Chem.* **2007**, *79*, 1708-1713.
- (38) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. *Mol. Cell. Proteomics*. **2007**, *6*, 1942-1951.
- (39) Chalkley, R. J.; Medzihradszky, K. F.; Lynn, A. J.; Baker, P. R.; Burlingame, A. L. *Anal. Chem.* **2010**, *82*, 579-584.
- (40) Sun, R. X.; Dong, M. Q.; Song, C. Q.; Chi, H.; Yang, B.; Xiu, L. Y.; Tao, L.; Jing, Z. Y.; Liu, C.; Wang, L. H.; Fu, Y.; He, S. M. *J. Proteome Res.* **2010**, *9*, 6354-6367.

Table S1. GPD scoring list of model glycopeptides based on the training dataset.

Test	Charge	Candidate ^A	<i>m/z</i> ^B	Glycopeptide Composition	Score ^C
1	3+	A	1315.8759	LCPDCPLLAPLNDSR+ [Hex]5[HexNAc]4[Neu5Ac]2 ^D	51.8
	3+	B	1315.5880	ELQTNALVCVENTTDLASILIK+ [Hex]5[HexNAc]3[SO3]1	4.6
	3+	C	1315.5915	QTTRVNAESTENSSLLTIK+ [Hex]5[HexNAc]5	4.4
	3+	D	1315.9330	TYFTNNVATLVFNKVNINDSGEYTCK+ [Hex]2[HexNAc]3	3.8
	3+	E	1315.8966	LVINRTHASDEGPYK+ [Hex]4[HexNAc]5[Neu5Ac]2	3.2
2	4+	A	921.1364	CGLVPVLAENY ^N K+[Hex]5[HexNAc]4[Neu5Ac]2	63.9
	4+	B	920.8886	VFAENETGLSRPR+[Hex]5[HexNAc]4[Neu5Ac]2	5.9
	4+	C	920.9131	V ^N KTP ^I ADLKFR+[Hex]6[HexNAc]5[Neu5Ac]1	4.4
	4+	D	920.6646	YDVPGPPL ^N VTITDVNR+[Hex]4[HexNAc]5[Fuc]1	3.1
	4+	E	920.8855	^N VTFTSVIR+[Hex]7[HexNAc]6[Neu5Ac]1	1.4
3	5+	A	1147.6926	CGLVPVLAENY ^N KSDNCEDTPEAGYFAVAVVK+ [Hex]5[HexNAc]4[Neu5Ac]2	37.7
	5+	B	1147.6768	I ^N ETLELLSESPVYSTK+ [Hex]9[HexNAc]8[Fuc]1[Neu5Ac]2	6.4
	5+	C	1147.7283	EYTIVVKVLDTPGPPV ^N VTVKEISK+ [Hex]8[HexNAc]7[Neu5Ac]1	2.2

5+	D	1147.7257	YTVTATNSAGTATENLSVIVLEKPGPPVGPVR+	1.6
			[Hex]5[HexNAc]4[Neu5Ac]3	
5+	E	1147.6612	NLTEGEEYTFQVMAVNSAGR+	0.7
			[Hex]7[HexNAc]6[Fuc]2[Neu5Ac]3	
4	5+	1001.6634	VVHAVEVALATFNAESNGSYLQLVEISR+	50.1
			[Hex]6[HexNAc]5	
5+	B	1001.6063	CNYTNIQETYFEVTELTEDQR+	11.4
			[Hex]5[HexNAc]4[Fuc]1[Neu5Ac]2	
5+	C	1001.6212	ENCTISWENPLDNGGSEITNFIVEYR+	7.0
			[Hex]7[HexNAc]4	
5+	D	1001.6483	NVTVIEGESVTLECHISGYPSPTVTWYR+	3.9
			[Hex]4[HexNAc]5[Fuc]1	
5+	E	1001.5943	NAAGNFSEPSDSSGAITAR+	2.0
			[Hex]6[HexNAc]5[Neu5Ac]4	
5	6+	980.2628	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR+	59.6
			[Hex]5[HexNAc]4[Neu5Ac]2	
6+	B	980.2327	MCLLNWSDPEDDGGSEITGFIIR+	13.9
			[Hex]6[HexNAc]7[Fuc]1[Neu5Ac]2	
6+	C	980.2480	YSLKATNGSGQATSTAELLVK+	3.9
			[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]3	
6+	D	980.2624	WTTDGSEIKTDEHYTVETDNFSSVLTIKNCLR+	2.9
			[Hex]5[HexNAc]5[Neu5Ac]1	
6+	E	980.2648	SLQAVITNLTQGEEYLFRVVAVNEK+	2.0
			[Hex]6[HexNAc]5[Fuc]1[Neu5Ac]3	
6	5+	944.7851	QQQHLFGSNVTDCSGNFCLFR+	44.2
			[Hex]5[HexNAc]4[Neu5Ac]2	
5+	B	944.6123	CYLAWNPPPLQDGGANISHYIIIEK+	7.6
			[Hex]5[HexNAc]4[Fuc]1[Neu5Ac]1	

7	5+	C	944.7861	NNTLVLR+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]4	4.5
	5+	D	944.8233	NEIASHENVTLPPTPLDPVLAAGR+ [Hex]5[HexNAc]4[Neu5Ac]2	4.1
	5+	E	944.8090	LLQNSENTIENTEHYTHLVMK+ [Hex]6[HexNAc]5[Fuc]1[NeuGc]3	2.6
	5+	F	944.6361	VADPIERPSPPVNTSSDQTQSSVQLK+ [Hex]5[HexNAc]5	1.9
	3+	A	1243.5230	LCPDCPLLAPLNDNR+ [Hex]6[HexNAc]5	38.5
	3+	B	1243.5142	NETVIEKPTDALQITK+ [Hex]5[HexNAc]4[Fuc]1[SO3]2	4.5
8	3+	C	1243.5289	MSDAGKYTVVAGGNVSTAK+ [Hex]6[HexNAc]3[Neu5Ac]1	3.9
	3+	D	1243.5074	DSGYYSLTAENSSGTDQK+ [Hex]3[HexNAc]6	1.9
	3+	E	1243.5209	AQIEVTSSFTMLVIDNVTR+ [Hex]5[HexNAc]3[Fuc]1[SO3]1	0
	3+	A	1227.8462	CGLVPVLAENYNK+[Hex]5[HexNAc]4[Neu5Ac]2	64.2
	3+	B	1227.5311	NNTLVLRK+[Hex]5[HexNAc]4[Neu5Ac]3	18.1
9	3+	C	1227.5157	VFAENETGLSRPR+[Hex]5[HexNAc]4[Neu5Ac]2	3.9
	3+	D	1227.5484	VNKTPIADLKFR+[Hex]6[HexNAc]5[Neu5Ac]1	3.2
	3+	E	1227.5116	NVTFTSVIR+[Hex]7[HexNAc]6[Neu5Ac]1	0
	6+	A	944.0864	RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR+ [Hex]6[HexNAc]5	46.8

6+	B	944.0688	GQVDLVDTMAFLVIPNSTRDDSGK+	7.6	
			[Hex]7[HexNAc]6[Fuc]1[Neu5Ac]2		
	C	944.1929	DSGDYTITAENSSGSK+	3.5	
			[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]4		
	D	944.0883	TLKNLTVTETQDAVFTVELTHPNVK+	0.9	
			[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]4		
6+	E	944.0603	HILVINDSQFDDEGVYTAEEVGK+	0.8	
			[Hex]7[HexNAc]6[Fuc]1[Neu5Ac]2		
10	4+	A	1251.8274	VVHAVEVALATFNAESNGSYLQLVEISR+	26.9
				[Hex]6[HexNAc]5	
	4+	B	1251.5665	ELQTNALVCVENTTDLASILIKDADR+	6.7
				[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]1	
	4+	C	1251.7483	DDEGKYTFYAGENITSGK+	5.6
			[Hex]8[HexNAc]7[Neu5Ac]1		
4+	D	1251.7746	ENCTISWENPLDNGGSEITNFIVEYR+	3.2	
			[Hex]7[HexNAc]4		
4+	E	1251.7946	DNKEIRPGGNYTITCVGNTPHLR+	0	
			[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]2		
11	4+	A	1180.7296	QQQHLFGSNVTDCSGNFCLFR+	35.6
				[Hex]5[HexNAc]4[Neu5Ac]2	
	4+	B	1180.5432	VADPIERPSPPVNLTSSDQTQSSVQLK+	8.4
				[Hex]5[HexNAc]5	
	4+	C	1180.7607	GQVDLVDTMAFLVIPNSTR+	6.8
			[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]3		
4+	D	1180.7594	LLQNSENITIENTEHYTHLVMK+	5.0	
			[Hex]6[HexNAc]5[Fuc]1[NeuGc]3		
4+	E	1180.5136	CYLAWNPPPLQDGGANISHYIIIEK+	0	
			[Hex]5[HexNAc]4[Fuc]1[Neu5Ac]1		

12	3+	A	1218.8441	LCPDCPLLAPL ^N DSR+	29.7
				[Hex]5[HexNAc]4[Neu5Ac]3	
	3+	B	1218.8154	YTLTVE ^N NSGSK+	5.5
				[Hex]6[HexNAc]3[Fuc]1[NeuGc]2	
	3+	C	1218.8652	STFEISSVQASDEG ^N YSVVVENSEGK+	3.9
				[Hex]3[HexNAc]2	
13	3+	D	1218.8732	YDVPGPPL ^N VTITDVNR+	0
				[Hex]6[HexNAc]4	
	3+	E	1218.8536	I ^N GSEPLQVSWYK+	0
				[Hex]6[HexNAc]5[Fuc]1	
	5+	A	1307.3594	RPTGEVYDIEIDTLETTCHVLDPTPLAN ^N CSVR+	25.1
				[Hex]6[HexNAc]5[Neu5Ac]3	
14	5+	B	1307.1487	WLKC ^N YTIVSDNFFT ^N VTALSEGDTYEFR+	8.7
				[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]1	
	5+	C	1307.3740	DDANLQTSFVHN ^N VATLQILQTDQSHIGQY ^N CSAS	7.7
				NPLGTASSSAK+[Hex]4[HexNAc]5[SO3]1	
	5+	D	1307.5418	DSV ^N L ^N LTWTEPASDGGSKITNYIVEK+	6.8
				[Hex]7[HexNAc]6[Fuc]2[Neu5Ac]4	
14	5+	E	1307.3548	YRMTFIDSVAVIQMN ^N LSTEDSGDFICEAQNPAG	5.5
				STSCSTK+[Hex]6[HexNAc]5	
	5+	A	1176.3206	VVHAVEVALATFNAES ^N GSYLQLVEISR+	20.4
				[Hex]6[HexNAc]5[Neu5Ac]3	
	5+	B	1176.2963	DSV ^N L ^N LTWTEPASDGGSKITNYIVEK+	8.8
				[Hex]6[HexNAc]5[Fuc]2[Neu5Ac]3	
14	5+	C	1176.1162	SLQAVIT ^N L ^N TQGEEYLFRVVAVNEK+	4.9
				[Hex]6[HexNAc]5[Fuc]1[NeuGc]3	
	5+	D	1176.2768	CHYMTIH ^N VTPDDEGVYSVIAR+	4.4
				[Hex]8[HexNAc]7[Neu5Ac]2	
	5+	E	1176.2956	LMLQ ^N ISPSDAGEYTA ^N VVGQLECKAK+	

				[Hex]6[HexNAc]5[Fuc]1[NeuGc]3	2.1
15	3+	A	1164.5106	KLCPDCPLLAPL ^N DSR+	30.7
				[Hex]5[HexNAc]4	
	3+	B	1164.5141	LLERPPEFTLPLY ^N K+	3.9
				[Hex]6[HexNAc]3[SO3]1	
	3+	C	1164.4985	VSVESSAV ^N TTTLIVYDCQ+	2.9
				[Hex]3[HexNAc]4[SO3]1	
	3+	D	1164.5183	V ^N KSLLNALK+	0
				[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]2	
	3+	E	1164.4934	ARTEIISTD ^N HHTLLTVK+	0
				[Hex]5[HexNAc]3[SO3]2	
16	3+	A	1286.2214	KLCPDCPLLAPL ^N DSR+	47.7
				[Hex]6[HexNAc]5	
	3+	B	1286.2235	VSVESSAV ^N TTTLIVYDCQK+	18.1
				[Hex]7[HexNAc]3	
	3+	C	1286.1767	WVRVNK+	5.8
				[Hex]6[HexNAc]5[Fuc]1[NeuGc]3	
	3+	D	1286.2242	WNEPKD ^N GSPILGYWLEK+	2.5
				[Hex]5[HexNAc]3[Neu5Ac]1	
	3+	E	1286.2220	NSILWVKL ^N K+	0
				[Hex]5[HexNAc]4[Fuc]1[Neu5Ac]3	
17	4+	A	964.9179	KLCPDCPLLAPL ^N DSR+	33.2
				[Hex]6[HexNAc]5	
	4+	B	964.9183	NSILWVKL ^N K+	7.2
				[Hex]5[HexNAc]4[Fuc]1[Neu5Ac]3	
	4+	C	964.9434	I ^N GSEPLQVSWYKDGVLK+	7.1
				[Hex]5[HexNAc]3[Neu5Ac]1	
	4+	D	964.9061	FDSGRY ^N LTLNNSGSK+	6.6
				[Hex]4[HexNAc]5[Neu5Ac]1	

	4+	E	964.8792	SNCTVSVHVSDR+	5.7
				[Hex]5[HexNAc]4[Neu5Ac]3	
	4+	F	964.7254	LFIAYQGRPTPTAVWSKPDSNLSLR	2.5
				[Hex]3[HexNAc]2[Fuc]1	
	4+	G	964.8917	VQIEKGVNYTQLSIDNCDR	2.2
				[Hex]3[HexNAc]4[Fuc]1[SO3]2	
18	3+	A	645.6193	SRNLTK+[Hex]5[HexNAc]2	25.4
	3+	B	645.6249	WVRHNK+[Hex]3[HexNAc]3	11.8
	3+	C	645.5645	NR+[Hex]5[HexNAc]3[Fuc]1[SO3]1	8.5
	3+	D	645.6053	WVRVNVK+[Hex]4[HexNAc]2[PO3]1	4.4
	3+	E	645.6123	AYANVSSK+[Hex]3[HexNAc]3	4.3
19	3+	A	681.9778	SRNLTKDR+[Hex]4[HexNAc]2	53.6
	3+	B	681.9791	YTLTLENSSGTK+[Hex]2[HexNAc]2	5.0
	3+	C	681.9824	NGTEILKSK+[Hex]4[HexNAc]2	3.6
	3+	D	681.9882	LNDTNTNSSSGRMIC+[Hex]2	2.3
	3+	E	681.9154	HNK+[Hex]5[HexNAc]3[Fuc]1[SO3]1	0
20	3+	A	735.9954	SRNLTKDR+[Hex]5[HexNAc]2	41.4
	3+	B	736.0000	NGTEILKSK+[Hex]5[HexNAc]2	6.3
	3+	C	735.9967	YTLTLENSSGTK+[Hex]3[HexNAc]2	4.9
	3+	D	736.0499	NSILWQKVNTSPISGR+[HexNAc]2	2.0

	3+	E	736.0130	VFAENETGLSRPR+[Hex]2[HexNAc]2	0
21	3+	A	790.0130	SRNLTKDR+[Hex]6[HexNAc]2	28.6
	3+	B	790.0143	YTLTLENSSGTK+[Hex]4[HexNAc]2	6.4
	3+	C	790.0055	VNVSSSK+[Hex]3[HexNAc]5[Fuc]1	4.9
	3+	D	790.0218	NGINVTSPQR+[Hex]2[HexNAc]4[Fuc]1	3.7
	3+	E	790.0176	NGTEILKSK+[Hex]6[HexNAc]2	2.1
22	4+	A	1160.5444	VVHAVEVALATFNAESNGSYQLVEISR+ [Hex]5[HexNAc]4	21.1
	4+	B	1160.4916	ENCTISWENPLDNGGSEITNFIVEYR+ [Hex]6[HexNAc]3	5.2
	4+	C	1160.5296	FGPPGPPEKPEVSNVTKNATVSWK+ [Hex]5[HexNAc]5[Fuc]1	4.7
	4+	D	1160.5186	TDTMRLLERPPEFTLPLYNK+ [Hex]5[HexNAc]4[Neu5Ac]2	3.2
	4+	E	1160.5255	NVTVIEGESVTLECHISGYPSPTVTWYR + [Hex]3[HexNAc]4[Fuc]1	1.2
23	4+	A	865.3685	WTNDLGSNMTIGAVNSR+ [Hex]5[HexNAc]4	39.9
	4+	B	865.3971	INVTDSLDTLTLISK+ [Hex]5[HexNAc]5	17.6
	4+	C	865.3582	DSVNLTWTEPASDGGSK+ [Hex]4[HexNAc]3[Fuc]1[Neu5Ac]1	9.1
	4+	D	865.3685	SNVAGINMTSGLDNTWR+ [Hex]5[HexNAc]4	4.2
	4+	E	865.3817	VNAESTENNSLLTIK+ [Hex]5[HexNAc]5	4.1

24	3+	A	1126.1378	WTNDLGS ^N MTIGAVNSR+	42.7
				[Hex]7[HexNAc]2	
	3+	B	1126.1174	NAAG ^N FSEPSDSSGAITAR+	12.8
				[Hex]3[HexNAc]4[Fuc]1[SO3]1	
	3+	C	1126.1554	VNAESTEN ^S NSLLTIK+	9.2
				[Hex]7[HexNAc]3	
	3+	D	1126.1623	WEPLDDGGSEI ^N YTLEK+	3.5
				[Hex]4[HexNAc]2[Fuc]1	
	3+	E	1126.1759	I ^N VTDSLDTL ^T SLIK+	2.0
				[Hex]7[HexNAc]3	
25	4+	A	1068.2094	LVPVPIT ^N ATLDR+	31.7
				[Hex]6[HexNAc]5[Neu5Ac]3	
	4+	B	1068.2004	LEN ^S SGSKSAFVTVK+	10.8
				[Hex]6[HexNAc]5[Fuc]1[Neu5Ac]2	
	4+	C	1068.1833	SDSGKYCVVVE ^N STGSR+	10.7
				[Hex]6[HexNAc]5[Fuc]1[Neu5Ac]1	
	4+	D	1068.2122	HILELS ^N LTIQDR+	4.9
				[Hex]8[HexNAc]7	
	4+	E	1068.2007	LVI ^N RTHASDEGPYK+	4.0
				[Hex]6[HexNAc]5[Neu5Ac]2	
26	4+	A	1140.2318	CANLVPVPIT ^N ATLDR+	20.2
				[Hex]6[HexNAc]5[Neu5Ac]3 ^E	
	4+	B	1140.2190	D ^N GSPILGYWLEKR+	5.4
				[Hex]7[HexNAc]5[Fuc]1[NeuGc]2	
	4+	C	1140.2318	MSDAGKYTVVAG ^N VSTAK+	5.1
				[Hex]7[HexNAc]7[Fuc]1	
	4+	D	1140.2110	^N LTEGEEYTFQVMAVNSAGR+	4.1
				[Hex]6[HexNAc]3[Fuc]1[NeuGc]2	
	4+	E	1140.2453	A ^N DTLVRSTEYPCAGLVEGLEYSFR+	0
				[Hex]5[HexNAc]3[Neu5Ac]1	
27	4+	A	1150.9832	CANLVPVPIT ^N ATLDR+	

				[Hex]6[HexNAc]5[Neu5Ac]3 ^F	18.7
	4+	B	1151.0178	QTTRINVTDSLDTTSLIK+ [Hex]6[HexNAc]6[Neu5Ac]1	5.5
	4+	C	1150.9942	YNLTLENSGSKTAFVNVR+ [Hex]9[HexNAc]5	5.1
	4+	D	1150.9658	NVDSVVNGTCRLDCK+ [Hex]8[HexNAc]7[Fuc]1	4.6
	4+	E	1150.9888	NDGRCHYMTIHNVTDPDEGVYSVIAR+ [Hex]6[HexNAc]3	2.4
28	4+	A	1180.7236	QDQCIYNTTYLNVQR+ [Hex]6[HexNAc]5[Neu5Ac]3 ^E	30.4
	4+	B	1180.7387	WVRCNFTDVSECQYTVTGLSPGDR+ [Hex]6[HexNAc]3[Neu5Ac]1	12.2
	4+	C	1180.7347	NSILWTKV ^N K+ [Hex]7[HexNAc]6[Neu5Ac]4	7.2
	4+	D	1180.7722	NSSGHAQGS ^A IVNVLDRPGPCQNLK+ [Hex]4[HexNAc]5[Fuc]1[Neu5Ac]1	3.2
	4+	E	1180.7607	GQVDLVDTMAFLVIP ^N STR+ [Hex]7[HexNAc]6[Neu5Ac]1	0
29	5+	A	1166.9045	SVQEIQATFFYFTP ^N KTEDTIFLR+ [Hex]7[HexNAc]6[Neu5Ac]2	31.7
	5+	B	1166.8805	ENCTISWENPLDNGGSEITNFIVEYR+ [Hex]6[HexNAc]6[Neu5Ac]2	6.2
	5+	C	1166.9297	FGISDHIDSACVTVKLPYTTGPPSTPWVT ^N VTR [Hex]5[HexNAc]5[Neu5Ac]1	5.6
	5+	D	1166.9224	SLQAVIT ^N LTQGEEYLFRVVAVNEK+ [Hex]8[HexNAc]7[Neu5Ac]1	0
	5+	E	1166.8782	NDTGKYILTIENG ^V GEPK+ [Hex]8[HexNAc]7[Neu5Ac]4	0
30	4+	A	1154.4871	CANLVPVPIT ^N ATLDR+ [Hex]6[HexNAc]5[Neu5Ac]3	18.3

4+	B	1154.4624	IV ^N LTENAGYYFR+ [Hex]6[HexNAc]5[Fuc]1[NeuGc]3	5.3
4+	C	1154.4972	QTTRVNAESTEN ^N SLLTIK+ [Hex]5[HexNAc]4[Neu5Ac]3	3.9
4+	D	1154.4950	IRDAHLDDQANY ^N VSLTNHR+ [Hex]5[HexNAc]5[Fuc]1[Neu5Ac]1	3.4
4+	E	1154.4776	VETNC ^N LSVEKIK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]2	0

^A For every test listed here, candidate A is the actual glycopeptide composition corresponding to the ETD-MS/MS spectrum being analyzed.

^B The theoretical m/z values of all the candidates fall into the mass error range of 200 ppm from the monoisotopic masses recorded in full scan spectra (MS^1).

^C The final scores were calculated by GPD program based on the parameter values described in the paper (spectral pre-processing: 5 tallest peaks preserved from each unit interval in region 1, 3 peaks from each unit in region 2, the intensity of remaining peaks in region 1 amplified by a factor of 5). The MS/MS scan range was from 50-2000 Da.

^D The glycosylated asparagine is marked as red N's in each composition.

^E Cysteine is unalkylated in the glycopeptide composition.

^F Carbamylation occurs on the *N*-terminus of the glycopeptide.

Table S2. GPD scoring list of glycopeptides from HIV Env Glycoprotein based on the validation dataset.

Test	Charge	Candidate ^A	m/z ^B	Glycopeptide Composition	Score ^C
1	3+	A	1002.1186	SVEIVCTRPN ^N NTRK+[Hex]5[HexNAc]2	34.0
	3+	B	1002.1186	KRTN ^N NPRTCIEVS+[Hex]5[HexNAc]2	9.5
	3+	C	1002.1467	YILTVEN ^N SSGSKSAFVNVR+[Hex]2[HexNAc]3	14.2
	3+	D	1002.1291	NNLPISISS ^N VSISR+[Hex]4[HexNAc]3[Fuc]1	6.1
	3+	E	1002.1113	IV ^N LTENAGYYFR+[Hex]3[HexNAc]4[Fuc]1	3.7

	3+	F	1002.0852	INTSAFELNER+[Hex]5[HexNAc]3[Neu5Ac]1	3.6
2	3+	A	1078.1555	SVEIVCTRPNNTRK+[Hex]3[HexNAc]4[Fuc]1	20.7
	3+	B	1078.1555	KRTNNPRTCIEVS+[Hex]3[HexNAc]4[Fuc]1	4.4
	3+	C	1078.1460	VNKTPIADLK+[Hex]6[HexNAc]5[Fuc]1	5.2
	3+	D	1078.1390	WLEVINITK+[Hex]5[HexNAc]5[Neu5Ac]1	0
	3+	E	1078.1460	LNKTPIPQTK+[Hex]7[HexNAc]4[Fuc]1	0
	3+	F	1078.1096	YTIEAENQSGKK+[Hex]9[HexNAc]2	0
3	3+	A	959.4201	SVEIVCTRPNNTR+[Hex]5[HexNAc]2	31.1
	3+	B	959.4201	RTNNPRTCIEVS+[Hex]5[HexNAc]2	9.6
	3+	C	959.4173	NSILWVKLNK+[Hex]6[HexNAc]3[SO3]1	10.7
	3+	D	959.3991	QNAVTVQGLIQGK+[Hex]7[HexNAc]2[PO3]1	3.2
	3+	E	959.4465	WNEPKDNGSPILGYWLEK+[Hex]2[HexNAc]2	2.4
	3+	F	959.4036	WVPVNVK+[Hex]6[HexNAc]5[Fuc]1	0
4	3+	A	1211.1855	SVEIVCTRPNNTR+ [Hex]5[HexNAc]5[Fuc]1	18.9
	3+	B	1211.1855	RTNNPRTCIEVS+ [Hex]5[HexNAc]5[Fuc]1	7.9
	3+	C	1211.1902	CHYMTIHNVTPDDEGVYSVIAR+ [Hex]4[HexNAc]2	5.4
	3+	D	1211.1769	HSTRWVPVNVK+	0

				[Hex]5[HexNAc]5[Neu5Ac]2	
5	3+	E	1211.2133	VVGPIRFTNITGEK+	0
				[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]1	
	3+	F	1211.1727	DAHLDDQANYNVSLTNHR+	0
				[Hex]4[HexNAc]3[Neu5Ac]1	
	3+	G	1211.1538	VFAENETGLSRPR+	0
				[Hex]6[HexNAc]4[Neu5Ac]1[SO3]1	
	3+	A	1089.4748	SVEIVCTRPNNNTR+[Hex]4[HexNAc]4[Fuc]1	22.6
	3+	B	1089.4748	RTNNNPRTCIEVS+[Hex]4[HexNAc]4[Fuc]1	5.3
	3+	C	1089.5101	NIEKVEVEAVNITQEPR+[Hex]3[HexNAc]4	6.0
	3+	D	1089.4795	TEIISTDNHTLLTVK+[Hex]3[HexNAc]5[SO3]1	0
	3+	E	1089.4565	LNGSAPIQVCWYR+[Hex]8[HexNAc]2	0
	3+	F	1089.4690	LNKTPIQDTK+[Hex]8[HexNAc]4	0
	3+	G	1089.4937	VLDSPSAPVNLITIR+[Hex]6[HexNAc]4	0
6	3+	A	1035.4572	SVEIVCTRPNNNTR+[Hex]3[HexNAc]4[Fuc]1	23.8
	3+	B	1035.4572	RTNNNPRTCIEVS+[Hex]3[HexNAc]4[Fuc]1	4.6
	3+	C	1035.4761	VLDSPSAPVNLITIR+[Hex]5[HexNAc]4	6.8
	3+	D	1035.4246	LNGSAPIQVCWYR+[Hex]4[HexNAc]4[SO3]1	4.6
	3+	E	1035.4761	TEIISTDNHTLLTVK+[Hex]5[HexNAc]3	3.3
	3+	F	1035.4217	VNINDSGEYTCK+[Hex]3[HexNAc]6	2.9
	3+	A	967.7641	SVEIVCTRPNNNTR+[Hex]3[HexNAc]3[Fuc]1	18.4

	3+	B	967.7641	RTN N NPRTCVEVS+[Hex]3[HexNAc]3[Fuc]1	10.9
	3+	C	967.7281	N VTFTSVIR+[Hex]6[HexNAc]4[SO3]1	4.1
	3+	D	967.7830	VLDSPSAPV N LTIR+[Hex]5[HexNAc]3	0
	3+	E	967.7687	TEIISTD N HTLLTVK+[Hex]2[HexNAc]4[SO3]1	0
	3+	F	967.7143	VDMNDAG N FTCR+[Hex]3[HexNAc]5	0
	3+				
8	3+	A	934.7438	TIIVHL N K+[Hex]9[HexNAc]2	36.6
	3+	B	934.7438	K N LHVIIT+[Hex]9[HexNAc]2	12.9
	3+	C	934.7347	WLEVIN I TK+[Hex]4[HexNAc]4[Fuc]1[SO3]1	5.9
	3+	D	934.7128	NVDSVV N GTCTCR+[Hex]6[HexNAc]3	4.9
	3+	E	967.7597	V N KSLLNALK+[Hex]8[HexNAc]2	3.9
	3+	F	934.7326	DSV N LTWTEPASDGGSK+[Hex]3[HexNAc]2[Fuc]1	0
9	3+	A	880.7262	TIIVHL N K+[Hex]8[HexNAc]2	24.9
	3+	B	880.7262	K N LHVIIT+[Hex]8[HexNAc]2	5.5
	3+	C	880.7279	V N KSLLNALK+[Hex]4[HexNAc]4[SO3]1	11.8
	3+	D	880.7071	WVPV N K+[Hex]4[HexNAc]4[Fuc]1[Neu5Ac]1	0
	3+	E	880.7150	DSV N LTWTEPASDGGSK+[Hex]2[HexNAc]2[Fuc]1	0
	3+	F	880.7171	WLEVIN I TK+[Hex]3[HexNAc]4[Fuc]1[SO3]1	0
10	3+	A	826.7086	TIIVHL N K+[Hex]7[HexNAc]2	33.7

	3+	B	826.7086	K N LHVIIT+[Hex]7[HexNAc]2	6.8
	3+	C	826.7103	V N KSLLNALK+[Hex]3[HexNAc]4[SO3]1	7.3
	3+	D	826.6628	QRLV I N R + [Hex]5[HexNAc]3[SO3]2	0
	3+	E	826.6776	NVDSVV N GTCR+[Hex]4[HexNAc]3	0
	3+	F	826.6995	WLEVI N ITK+[Hex]2[HexNAc]4[Fuc]1[SO3]1	0
11	3+	A	727.0173	TIIVHL N K+[Hex]3[HexNAc]3[Fuc]1	35.8
	3+	B	727.0173	K N LHVIIT+[Hex]3[HexNAc]3[Fuc]1	3.9
	3+	C	726.9892	LCN N KTFNGTGPCK +[Hex]1[HexNAc]2	5.6
	3+	D	727.0248	N ASGSAKAEIKVK+[Hex]2[HexNAc]2[Fuc]1	2.6
	3+	E	727.0222	EINCTRP N NTRKR+[HexNAc]2	2.2
	3+	F	726.9331	N R+[Hex]4[HexNAc]5[Fuc]1[SO3]1	0
12	5+	A	1058.8502	DGGED N KTEEIFRPGGGNMK+ [Hex]7[HexNAc]7[Fuc]2[Neu5Ac]1	26.3
	5+	B	1058.8502	KMNGGGPRFIEETK N DEGGD+ [Hex]7[HexNAc]7[Fuc]2[Neu5Ac]1	7.5
	5+	C	1058.9017	AIARTAV N ISPPSESDPVTILAENVPPR+ [Hex]6[HexNAc]5[Neu5Ac]1	4.9
	5+	D	1058.8625	EPPSFVQKPDPM D VL T GT N VTFTSIVK+ [Hex]5[HexNAc]5[Fuc]1[Neu5Ac]1[SO3]1	4.9
	5+	E	1058.8429	MCLL N WSDPEDDGGSEITGFI E RK+ [Hex]5[HexNAc]5[Neu5Ac]2	4.1
	5+	F	1058.8873	YTVTATNSAGTATE N LSVIVLEKPGPPVGPVR+ [Hex]5[HexNAc]5[Fuc]1[SO3]1	3.6

	5+	G	1058.8535	MVDRLFPPGPPEKPEVS N VTK+ [Hex]8[HexNAc]7[Neu5Ac]1	0
13	4+	A	1381.5472	DGGED N KTEEIFRPGGGNMK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]3	52.9
	4+	B	1381.5472	KM N GGGPRFIEETKNDEGGD+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]3	8.3
	4+	C	1381.6695	LPFNTYSIQAGEDLKIEIPVIGRPRP N ISWVK+ [Hex]6[HexNAc]3[Neu5Ac]1	4.1
	4+	D	1381.5870	ASFT N VTETQFIISGLTQNSQYEFR+ [Hex]5[HexNAc]4[Fuc]1[Neu5Ac]3	0
	4+	E	1381.5888	CYLAWNPPPLQDGGAN I SHYIIIEK+ [Hex]8[HexNAc]7[Fuc]1	0
	4+	F	1381.6195	KPKDVTALE N ATVAFEVSVSHDTVPVK+ [Hex]5[HexNAc]4[Fuc]1[Neu5Ac]3	0
	4+	G	1381.5836	VSVESSAV N TTTLIVYDCQKSDAGK+ [Hex]8[HexNAc]6[Fuc]1[Neu5Ac]1	0
	4+	H	1381.6286	IEIPVIGRPRP N ISWVKDGEPLK+ [Hex]7[HexNAc]5[Fuc]1[NeuGc]2	0
14	4+	A	1235.9996	DGGED N KTEEIFRPGGGNMK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]1	24.5
	4+	B	1235.9996	KMNGGGPRFIEETK N DEGGD+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]1	7.1
	4+	C	1236.0190	WEPLDDGGSEI N YTLEK+ [Hex]8[HexNAc]5[Fuc]1[NeuGc]1	11.0
	4+	D	1236.0058	SVSLEVNNLEEDTAN Y TCK+ [Hex]5[HexNAc]4[Fuc]1[Neu5Ac]3	7.4
	4+	E	1236.0393	ASFT N VTETQFIISGLTQNSQYEFR+ [Hex]5[HexNAc]4[Fuc]1[Neu5Ac]1	0
	4+	F	1236.0450	ELQTNALVCVEN T TDLASILIK+ [Hex]5[HexNAc]4[Neu5Ac]3	0
15	4+	A	1173.4824	DGGED N KTEEIFRPGGGNMK+ [Hex]6[HexNAc]7[Fuc]1	23.4

16	4+	B	1173.4824	KMNGGGPRFIEETK ^N DEGGD+	5.1
				[Hex]6[HexNAc]7[Fuc]1	
	4+	C	1173.4597	STFEISSVQASDEG ^N YSVVVENSEGK+	9.7
				[Hex]5[HexNAc]4[Fuc]1[SO3]2	
	4+	D	1173.4886	SVSLEVNNLELED ^N TANYTCK+	5.4
				[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]2	
16	4+	E	1173.4538	C ^N YTIVSDNFFVTALSEGDTYEF ^R +	3.4
				[Hex]6[HexNAc]3[SO3]2	
	4+	F	1173.4508	^N DSELHESWK+	2.9
				[Hex]8[HexNAc]7[Fuc]1[Neu5Ac]2	
	4+	A	1122.7125	DGGED ^N KTEEIFRPGGGNMK+	20.9
				[Hex]6[HexNAc]6[Fuc]1	
16	4+	B	1122.7125	KM ^N GGGPRFIEETK ^N DEGGD+	11.6
				[Hex]6[HexNAc]6[Fuc]1	
	4+	C	1122.6837	EV ^N STHWSR+	11.3
				[Hex]7[HexNAc]6[Fuc]1[Neu5Ac]3	
	4+	D	1122.7111	LMLQ ^N ISPSDAGEYTAVVGQLECK+	11.2
				[Hex]3[HexNAc]6[SO3]2	
16	4+	E	1122.7549	EPPSFVQKPD ^N MDVLTGT ^N VTFTSIVK+	5.4
				[Hex]4[HexNAc]4[SO3]1	
	4+	F	1122.7145	WLEVI ^N ITK+	0
				[Hex]7[HexNAc]6[Fuc]1[Neu5Ac]3	
	4+	A	1154.9732	DGGED ^N KTEEIFRPGGGNMK+	22.9
				[Hex]5[HexNAc]6[Fuc]1[Neu5Ac]1	
17	4+	B	1154.9732	KMNGGGPRFIEETK ^N DEGGD+	10.3
				[Hex]5[HexNAc]6[Fuc]1[Neu5Ac]1	
	4+	C	1154.9824	LMLQ ^N ISPSDAGEYTAVVGQLECK+	6.9
				[Hex]5[HexNAc]4[Neu5Ac]1[SO3]1	
17	4+	D	1154.9821	WEPLDDGGSEI ^N YTLEK+	6.8
				[Hex]6[HexNAc]5[Fuc]1[NeuGc]1	

	4+	E	1154.9590	NAAGNFSEPSDSSGAITAR+ [Hex]8[HexNAc]5[Fuc]1[NeuGc]1	4.4
	4+	F	1154.9416	NDSELHESWK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]3	0.8
18	4+	A	1001.1728	DGGEDNKTEEIFRPGGGNMK+ [Hex]3[HexNAc]6[Fuc]1	22.6
	4+	B	1001.1728	KMNGGGPRFIEETKNDEGGD+ [Hex]3[HexNAc]6[Fuc]1	4.6
	4+	C	1001.1818	WEPLDDGGSEIINYTLEK+ [Hex]5[HexNAc]5	10.2
	4+	D	1001.2260	EPPSFVQKPDPM DVLGTNVTFTSIVK+ [Hex]4[HexNAc]2	7.0
	4+	E	1001.1532	DSVNLTWTEPASDGGSK+ [Hex]6[HexNAc]4[Fuc]1[NeuGc]1	5.3
	4+	F	1001.1483	SVSLEVNNLEEDTANYTCK+ [Hex]2[HexNAc]6[SO3]2	4.1
19	4+	A	950.4030	DGGEDNKTEEIFRPGGGNMK+ [Hex]3[HexNAc]5[Fuc]1	57.7
	4+	B	950.4030	KMNGGGPRFIEETKNDEGGD+ [Hex]3[HexNAc]5[Fuc]1	9.0
	4+	C	950.3714	NDSELHESWK+ [Hex]5[HexNAc]5[Fuc]1[Neu5Ac]2	16.8
	4+	D	950.3892	SVSLEVNNLEEDTANYTCK+ [Hex]5[HexNAc]3[SO3]1	5.3
	4+	E	950.4198	WEPLDDGGSEIINYTLEK+ [Hex]5[HexNAc]4	5.0
	4+	F	950.3864	YDVPGPPLNVTITDVNR+ [Hex]5[HexNAc]4[Fuc]1[SO3]2	5.0
20	3+	A	1199.1752	DGGEDNKTEEIFRPGGGNMK+	61.7

				[Hex]3[HexNAc]4[Fuc]1	
	3+	B	1199.1752	KMNGGGPRFIEETK N DEGGD+	13.3
				[Hex]3[HexNAc]4[Fuc]1	
	3+	C	1199.1564	NAAG N FSEPSDSSGAITAR+	7.1
				[Hex]7[HexNAc]3	
	3+	D	1199.1871	WEPLDDGGSEI N YTLEK+	6.4
				[Hex]5[HexNAc]3	
	3+	E	1199.1827	TCILEIL N STK+	5.4
				[Hex]4[HexNAc]6[Fuc]1[Neu5Ac]1	
	3+	F	1199.2023	L N GSAPIQVCWYRDGVLLR+	2.1
				[Hex]3[HexNAc]4[SO3]1	
21	4+	A	899.6332	DGGED N KTEEIFRPGGGNMK+	55.7
				[Hex]3[HexNAc]4[Fuc]1	
	4+	B	899.6332	KM N GGGPRFIEETKNDEGGD+	12.0
				[Hex]3[HexNAc]4[Fuc]1	
	4+	C	899.6433	V N KSLLNALK+	9.8
				[Hex]5[HexNAc]4[Neu5Ac]3	
	4+	D	899.6536	L N GSAPIQVCWYRDGVLLR+	5.4
				[Hex]3[HexNAc]4[SO3]1	
	4+	E	899.6486	YILTVEN S SGSKSAFVNVR+	3.1
				[Hex]3[HexNAc]4[Fuc]1[SO3]1	
	4+	F	899.6466	VLDTPGPPV N VTVKEISK+	2.9
				[Hex]2[HexNAc]6[SO3]2	
22	4+	A	848.8634	DGGED N KTEEIFRPGGGNMK+	55.4
				[Hex]3[HexNAc]3[Fuc]1	
	4+	B	848.8634	KM N GGGPRFIEETKNDEGGD+	4.8
				[Hex]3[HexNAc]3[Fuc]1	
	4+	C	848.8798	EFLC I NGSIHFQPLK+	7.6
				[Hex]3[HexNAc]4[Neu5Ac]1	

23	4+	D	848.8723	WEPPLDDGGSEIINYTLEK+	6.0
				[Hex]5[HexNAc]2	
	4+	E	848.8367	LENSSGSK+	5.5
				[Hex]6[HexNAc]5[Fuc]2[Neu5Ac]1	
	4+	F	848.8895	NDTGKYILTIENGVEPK+	3.7
				[Hex]3[HexNAc]4[Fuc]1	
	3+	A	1285.1910	DGGEDNKTEEIFRPGGGNMK+	55.4
				[Hex]8[HexNAc]2	
	3+	B	1285.1910	KMNGGGPRFIEETKNDEGGD+	9.9
				[Hex]8[HexNAc]2	
	3+	C	1285.2049	CNYTNIQETYFEVTELTEDQR+	5.8
				[Hex]4[HexNAc]2[Fuc]1	
	3+	D	1285.1921	NVDSVVNGTCRLDCK+	4.6
				[Hex]5[HexNAc]5[Neu5Ac]1	
	3+	E	1285.2029	VENLTEGAIIYYFR+	2.6
				[Hex]6[HexNAc]5[Neu5Ac]1	
	3+	F	1285.2199	INETLELLSESPVYSTK+	1.8
				[Hex]3[HexNAc]6[Fuc]1[SO3]1	
24	3+	A	1123.1382	DGGEDNKTEEIFRPGGGNMK+	55.1
				[Hex]5[HexNAc]2	
	3+	B	1123.1382	KMNGGGPRFIEETKNDEGGD+	12.5
				[Hex]5[HexNAc]2	
	3+	C	1123.1219	ANHTPESCPETK+	10.3
				[Hex]3[HexNAc]6[Fuc]2	
	3+	D	1123.1814	INETLELLSESPVYSTK+	6.0
				[Hex]3[HexNAc]4[Fuc]1	
	3+	E	1123.1340	YCVVVENSTGSR+	5.7
				[Hex]3[HexNAc]6[Fuc]2	
	3+	F	1123.1457	TCILEILNSTK+	5.5
				[Hex]6[HexNAc]4[Neu5Ac]1	

25	4+	A	1066.4444	DGGEDNKTEEIFRPGGGNMKDNWR+	23.6
				[Hex]7[HexNAc]2	
	4+	B	1066.1444	RWNDKMNGGGPRFIEETKNDDEGGD+	14.9
				[Hex]7[HexNAc]2	
	4+	C	1066.4802	LSQTEPVTLIKDIENQTVLK+	8.6
				[Hex]5[HexNAc]4[Neu5Ac]1[SO3]1	
	4+	D	1066.4314	VQIEKGVNYTQLSIDNCDR+	7.7
				[Hex]3[HexNAc]6[Fuc]1[SO3]2	
	4+	E	1066.4696	VIARNAAGNFSEPSDSSGAITAR+	7.4
				[Hex]5[HexNAc]5[Fuc]1	
	4+	F	1066.4332	NVTFTSVIR+	7.2
				[Hex]7[HexNAc]6[Neu5Ac]3	
	4+	G	1066.4458	KDSGYYSLTAENSSGTDQK+	7.0
				[Hex]3[HexNAc]8	
	4+	H	1066.4335	DSVNLTWTEPASDGGSK+	3.6
				[Hex]7[HexNAc]6[Fuc]1	
	4+	I	1066.4567	SLQAVITNLTQGEEYLFR+	0
				[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]1[SO3]1	
26	4+	A	1093.2156	DGGEDNKTEEIFRPGGGNMKDNWR+	35.7
				[Hex]3[HexNAc]5[Fuc]1	
	4+	B	1093.2156	RWNDKMNGGGPRFIEETKNDDEGGD+	9.6
				[Hex]3[HexNAc]5[Fuc]1	
	4+	C	1093.2304	AQIEVTSSFTMLVIDNVTR+	10.9
				[Hex]4[HexNAc]5[Neu5Ac]2	
	4+	D	1093.1959	RPDYDGGSPNLSYHVER+	9.5
				[Hex]5[HexNAc]5[Neu5Ac]2	
	4+	E	1093.1872	YTFYAGENITSGK+	9.5
				[Hex]6[HexNAc]6[Fuc]1[Neu5Ac]2	
	4+	F	1093.2292	CYLAWNPPQLDGGANISHYIIIEK+	9.1
				[Hex]5[HexNAc]3[Neu5Ac]1	

27	4+	A	1042.4458	DGGEDNKTEEIFRPGGGNMKDNWR+	27.1
				[Hex]3[HexNAc]4[Fuc]1	
	4+	B	1042.4458	RWNDKMNGGGPRFIEETKNDDEGGD+	11.5
				[Hex]3[HexNAc]4[Fuc]1	
	4+	C	1042.4261	RPDYDGGSPNLSYHVER+	6.0
				[Hex]5[HexNAc]4[Neu5Ac]2	
	4+	D	1042.4716	NLTVTETQDAVFTVELTHPNVK+	5.7
				[Hex]5[HexNAc]3[Neu5Ac]1	
	4+	E	1042.4722	LLERPPEFTLPLYNK+	4.5
				[Hex]6[HexNAc]6[Fuc]1	
	4+	F	1042.4448	CHYMTIHNVTPDDEGVYSVIAR+	3.1
				[Hex]3[HexNAc]4[Neu5Ac]1	
28	3+	A	1124.4643	AMYAPPIAGNITCK+	39.0
				[Hex]9[HexNAc]2	
	3+	B	1124.4643	KCTINGAIPPAYMA+	5.1
				[Hex]9[HexNAc]2	
	3+	C	1124.5135	YDVPGPPLNVTITDVNR+	4.1
				[Hex]3[HexNAc]5	
	3+	D	1124.4939	INGSEPLQVSWYK+	4.1
				[Hex]3[HexNAc]6[Fuc]1	
	3+	E	1124.4498	VNINDSGEYTCK+	0
				[Hex]5[HexNAc]5[Fuc]1	
	3+	F	1124.4899	TEIISTDNHTLLTVK+	0
				[Hex]4[HexNAc]4[Fuc]1[SO3]1	
29	3+	A	1070.4467	AMYAPPIAGNITCK+	24.8
				[Hex]8[HexNAc]2	
	3+	B	1070.4467	KCTINGAIPPAYMA+	12.1
				[Hex]8[HexNAc]2	
	3+	C	1070.4742	GQVDLVDTMAFLVIPNSTR+	3.6

				[Hex]4[HexNAc]2[PO3]1	
	3+	D	1070.4465	WVRV N K+	0
				[Hex]5[HexNAc]5[Neu5Ac]2	
	3+	E	1070.4414	V N TSPISGR+	0
				[Hex]6[HexNAc]5[Neu5Ac]1	
	3+	F	1070.4350	L N GSAPIQVCWYR+	0
				[Hex]5[HexNAc]3[Fuc]1[SO3]1	
30	3+	A	1016.4291	AMYAPPIAG N ITCK+[Hex]7[HexNAc]2	36.7
	3+	B	1016.4291	KCTI N GAIPPAYMA+[Hex]7[HexNAc]2	5.5
	3+	C	1016.4326	YEITAA N SSGTTK+[Hex]3[HexNAc]6	14.2
	3+	D	1016.4322	DI E NQTVLK+[Hex]6[HexNAc]5	0
	3+	E	1016.4146	VNI N DSGEYTCK+[Hex]3[HexNAc]5[Fuc]1	0
	3+	F	1016.4690	VLDSPSAPVNL T IR+[Hex]5[HexNAc]3[Fuc]1	0
31	3+	A	962.4115	AMYAPPIAG N ITCK+[Hex]6[HexNAc]2	34.7
	3+	B	962.4115	KCTI N GAIPPAYMA+[Hex]6[HexNAc]2	12.3
	3+	C	962.3998	L N GSAPIQVCWYR+	3.1
				[Hex]3[HexNAc]3[Fuc]1[SO3]1	
	3+	D	962.4271	GVYTVEAK N ASGSAK+	0
				[Hex]4[HexNAc]3[Fuc]1	
	3+	E	962.4146	DI E NQTVLK+[Hex]5[HexNAc]5	0
	3+	F	962.4514	TEI I STD N HTLLTVK+[Hex]4[HexNAc]2[Fuc]1	0
32	3+	A	908.3939	AMYAPPIAG N ITCK+[Hex]5[HexNAc]2	33.0

	3+	B	908.3939	KCTINGAIPPAYMA+[Hex]5[HexNAc]2	0
	3+	C	908.3970	DIENQTVLK+[Hex]4[HexNAc]5	6.7
	3+	D	908.3599	SNCTVSVHVSDR+[Hex]2[HexNAc]4[Fuc]1[SO3]1	6.0
	3+	E	908.4091	IENTTTVLK+[Hex]3[HexNAc]6	6.0
	3+	F	908.4095	GVYTVEAKNASGSAK+[Hex]3[HexNAc]3[Fuc]1	3.5
33	3+	A	1119.8172	AMYAPPIAGNITCK+ [Hex]3[HexNAc]6[Fuc]1	26.4
	3+	B	1119.8172	KCTINGAIPPAYMA+ [Hex]3[HexNAc]6[Fuc]1	5.1
	3+	C	1119.7816	GVNYTQLSIDNCDR+ [Hex]8[HexNAc]2	10.1
	3+	D	1119.8641	LLQNSENITIENTEHYTHLVMK+ [Hex]2[HexNAc]2	7.6
	3+	E	1119.8043	YILTVENSSGSK+ [Hex]5[HexNAc]4[Fuc]1[Neu5Ac]1	0
	3+	F	1119.7975	SNCTVSVHVSDR+ [Hex]3[HexNAc]6[Fuc]2	0
34	5+	A	1161.4703	DGGEDNKTEEIFRPGGGNMKDNWR+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]2	35.7
	5+	B	1161.4703	RWNDKMNGGGPRFIEETKNDEGGD+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]2	5.4
	5+	C	1161.4728	NPYDPPGRCDPPVISNITK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]4	5.6
	5+	D	1161.4574	GVNYTQLSIDNCDRNDAGK+ [Hex]7[HexNAc]6[Fuc]1[Neu5Ac]4	4.7

35	5+	E	1161.5049	VADPIERPSPPV ^N LTSSDQTQSSVQLK+ [Hex]7[HexNAc]5[Fuc]1[NeuGc]2	4.4
	5+	F	1161.5202	VEHVKVTVLDPGPPGPVEIS ^N VSAEK+ [Hex]6[HexNAc]5[Fuc]1[Neu5Ac]3	0
	3+	A	1224.5229	LFNN ^N ATEDETITLPCR+ [Hex]4[HexNAc]5	43.3
	3+	B	1224.5229	RCPLTITETETAN ^N NNFL+ [Hex]4[HexNAc]5	0
	3+	C	1224.5002	DSGYYSLTAEN ^S SGTDTQK+ [Hex]3[HexNAc]5[Fuc]1	7.0
	3+	D	1224.5575	SLQAVIT ^N LTQGEEYLF+ [Hex]3[HexNAc]4[Neu5Ac]1	0
36	3+	E	1224.5209	NSVGKS ^N CTVSVHVSDR+ [Hex]5[HexNAc]5	0
	3+	F	1224.5366	LVIN ^R THASDEGPYK+ [Hex]5[HexNAc]5[Fuc]1	0
	3+	A	1247.3137	LTPLCVTLHCT ^N ATFK+[Hex]9[HexNAc]2	44.7
	3+	B	1247.3137	KFTAN ^T CHLTVCLPTL+[Hex]9[HexNAc]2	8.4
	3+	C	1247.2400	NSLLWVKAN ^D TLVR+[Hex]3[HexNAc]8	4.4
	3+	D	1247.4909	DTDQWYRVHT ^N ATIR+[Hex]3[HexNAc]6[SO3]2	3.9
37	3+	E	1247.1961	TDEHYTVETD ^N FSSVLTIK+[Hex]7[HexNAc]2	3.6
	3+	F	1247.2341	GQVDLVDTMAFLVIP ^N STR+[Hex]4[HexNAc]5	2.5
	3+	A	1193.2961	LTPLCVTLHCT ^N ATFK+[Hex]8[HexNAc]2	26.9
	3+	B	1193.2961	KFTAN ^T CHLTVCLPTL+[Hex]8[HexNAc]2	9.6
	3+	C	1193.1785	TDEHYTVETD ^N FSSVLTIK+[Hex]6[HexNAc]2	

					10.3
	3+	D	1193.4683	DSGDYTITAENSSGSK+[Hex]7[HexNAc]4	6.5
	3+	E	1193.1692	YILTVENSSGSK+[Hex]6[HexNAc]5[Fuc]2	5.1
	3+	F	1193.2165	GQVDLVDTMAFLVIPNSTR+[Hex]3[HexNAc]5	0
38	3+	A	1139.2785	LTPLCVTLHCTNATFK+ [Hex]7[HexNAc]2	20.1
	3+	B	1139.2785	KFTANTCHLTVCLPTL+ [Hex]7[HexNAc]2	7.1
	3+	C	1139.1763	VNKTIHDTQFK+ [Hex]5[HexNAc]5[Fuc]1	7.4
	3+	D	1139.1495	YILKLENSSGSK+ [Hex]3[HexNAc]6[Fuc]2[SO3]1	6.3
	3+	E	1139.4699	DTDQWYRVHTNATIR+ [Hex]4[HexNAc]4[SO3]1	5.7
	3+	F	1139.4507	DSGDYTITAENSSGSK+ [Hex]6[HexNAc]4	4.4
39	3+	A	1031.2433	LTPLCVTLHCTNATFK+[Hex]5[HexNAc]2	37.7
	3+	B	1031.2433	KFTANTCHLTVCLPTL+[Hex]5[HexNAc]2	10.0
	3+	C	1031.1411	VNKTIHDTQFK+[Hex]3[HexNAc]5[Fuc]1	6.6
	3+	D	1031.4155	DSGDYTITAENSSGSK+[Hex]4[HexNAc]4	6.1
	3+	E	1031.1091	VNRLNVTLK+[Hex]4[HexNAc]5[NeuNAc]1[SO3]1	0
	3+	F	1031.1257	TDEHYTVETDNFSSVLTIK+[Hex]3[HexNAc]2	0
40	3+	A	1242.6666	LTPLCVTLHCTNATFK+	

				[Hex]3[HexNAc]6[Fuc]1	40.5
3+	B	1242.6666	KFTAN ⁺ TCHLTVCLPTL+		16.7
				[Hex]3[HexNAc]6[Fuc]1	
3+	C	1242.6012	IEWSK ⁺ NETVIEKPTDALQITK+		6.9
				[Hex]2[HexNAc]4[Fuc]1	
3+	D	1242.5779	EYTIVVKVLDTPGPPV ⁺ NVTVK+		3.4
				[Hex]6[HexNAc]2[PO3]1	
3+	E	1242.8191	N ⁺ DSELHESWK+		3.2
				[Hex]6[HexNAc]6[Neu5Ac]1	
3+	F	1242.6012	IEIPVIGRPRP ⁺ ISWVK+		0
				[Hex]4[HexNAc]4[Neu5Ac]1	
41	3+	A	1174.9735	LTPLCVTLHCT ⁺ NATFK+	33.2
				[Hex]3[HexNAc]5[Fuc]1	
3+	B	1174.9735	KFTAN ⁺ TCHLTVCLPTL+		6.4
				[Hex]3[HexNAc]5[Fuc]1	
3+	C	1174.9080	IEIPVIGRPRP ⁺ ISWVK+		3.9
				[Hex]4[HexNAc]3[Neu5Ac]1	
3+	D	1174.8679	Q ⁺ NATVQGLIQGKAYFFR+		3.5
				[Hex]6[HexNAc]3	
3+	E	1174.8779	MVDRFGPPGPPEKPEVS ⁺ NVTK+		2.7
				[Hex]3[HexNAc]3[Fuc]1	
3+	F	1174.8476	I ⁺ NVTDSLDTLTL ⁺ SIK+		2.5
				[Hex]4[HexNAc]5[Fuc]1[SO3]1	
42	3+	A	1107.2804	LTPLCVTLHCT ⁺ NATFK+	29.7
				[Hex]3[HexNAc]4[Fuc]1	
3+	B	1107.2804	KFTAN ⁺ TCHLTVCLPTL+		9.3
				[Hex]3[HexNAc]4[Fuc]1	
3+	C	1107.2149	IEWSK ⁺ NETVIEKPTDALQITK+		4.1
				[Hex]2[HexNAc]2[Fuc]1	
3+	D	1107.2010	QTTRIN ⁺ VTDSDLDTLTL ⁺ SIK+		2.3

				[Hex]4[HexNAc]2[Fuc]1	
	3+	E	1107.1545	INVTDSLDTTSLIK+ [Hex]4[HexNAc]4[Fuc]1[SO3]1	2.3
	3+	F	1107.4395	LENSSGSK+ [Hex]7[HexNAc]6[Fuc]1	0
43	3+	A	984.1054	SNITGLLLVR+[Hex]9[HexNAc]2	38.0
	3+	B	984.1054	RVLLLGTTINS+[Hex]9[HexNAc]2	4.8
	3+	C	984.0779	DEIDAPNASLDPK+[Hex]5[HexNAc]3[Fuc]1	6.5
	3+	D	984.0879	KAYATITNCTK+[Hex]5[HexNAc]3[Fuc]1	4.8
	3+	E	984.0934	LNGSAPIQVCWYR+[Hex]3[HexNAc]3[Neu5Ac]1	3.9
	3+	F	984.0896	LNVTLK+[Hex]5[HexNAc]5[Fuc]1[NeuNAc]1	0
44	3+	A	930.0878	SNITGLLLVR+[Hex]8[HexNAc]2	42.9
	3+	B	930.0878	RVLLLGTTINS+[Hex]8[HexNAc]2	20.7
	3+	C	930.0720	LNVTLK+[Hex]4[HexNAc]5[Fuc]1[Neu5Ac]1	5.2
	3+	D	930.0552	NGINVTSPQR+[Hex]8[HexNAc]2	4.4
	3+	E	930.0725	NVTFTSVIR+[Hex]4[HexNAc]4[Neu5Ac]1	2.9
	3+	F	930.1013	TSVRLNWTKEHDGGAK+[Hex]3[HexNAc]2	0
45	3+	A	911.7452	SNITGLLLVR+[Hex]3[HexNAc]5[Fuc]1	48.2
	3+	B	911.7452	RVLLLGTTINS+[Hex]3[HexNAc]5[Fuc]1	8.0
	3+	C	911.7269	YTFYAGENITSGK+[Hex]2[HexNAc]4[Fuc]1	4.8

3+	D	911.7778	NETVIEKPTDALQITK+[Hex]2[HexNAc]3	2.4
3+	E	911.7141	NGTEILKSK+[Hex]4[HexNAc]5[SO3]1	0
3+	F	911.7326	NGINVTSPQR+[Hex]3[HexNAc]5[Fuc]1	0

^A For every test listed here, candidate A is the actual glycopeptide composition corresponding to the ETD-MS/MS spectrum being analyzed.

^B The theoretical m/z values of all the candidates fall into the mass error range of 200 ppm from the monoisotopic masses recorded in full scan spectra (MS¹).

^C The total scores were calculated by GPD program based on the parameters described in the paper (spectral pre-processing: 5 tallest peaks preserved from each unit interval in region 1, 3 peaks from each unit in region 2, the intensity of remaining peaks in region 1 amplified by a factor of 5). The MS/MS scan range was from 50-2000 Da.

CHAPTER III

Characterizing O-linked glycopeptides by electron transfer dissociation: fragmentation rules and applications in data analysis

This work has been published by the journal Analytical Chemistry, with reprint permission from the journal.

Studying protein *O*-glycosylation remains an analytical challenge. Different from *N*-linked glycans, the *O*-glycosylation site is not within a known consensus sequence. Additionally, *O*-glycans are heterogeneous with numerous potential modification sites. Electron transfer dissociation (ETD) is the method of choice in analyzing these glycopeptides since the glycan side chain is intact in ETD, and the glycosylation site can be localized on the basis of the *c* and *z* fragment ions. Nonetheless, new software is necessary for interpreting *O*-glycopeptide ETD spectra in order to expedite the analysis workflow. To address the urgent need, we studied the fragmentation of *O*-glycopeptides in ETD and found useful rules that facilitate their identification. By implementing the rules into an algorithm to score potential assignments against ETD-MS/MS data, we applied the method to glycopeptides generated from various *O*-glycosylated proteins including mucin, erythropoietin, fetuin and an HIV envelope protein, 1086.C gp120. The site-specific *O*-glycopeptide composition was correctly assigned in every case, proving the merits of our method in analyzing glycopeptide ETD data. The algorithm described herein can be easily incorporated into other automated glycomics tools.

3.1 Introduction

O-linked glycosylation typically occurs on serine and threonine residues in a protein, with the glycan portion bonded to a hydroxyl group on the amino acid's side-chain.¹ Mucin-type *O*-glycan is the most commonly seen *O*-glycosylation form; it contains an α -*N*-acetylgalactosamine (GalNAc) core structure.²⁻³ Recent studies indicate that aberrant mucin-type *O*-glycosylation on membrane proteins of tumor cells is closely related to cancer metastasis, and the tumor-specific glycosylation can be mimicked to develop glycoprotein-based cancer vaccine.⁴⁻⁵ It is thus a prerequisite to unravel the glycosylation profile on proteins.

However, *O*-linked glycopeptide analysis by mass spectrometry (MS) has long been a tedious task.^{2, 6} Unlike *N*-linked glycosylation, no consensus sequence is available to predict potential *O*-glycosylation sites; there are eight different basic structures in mucin-type GalNAc-linked glycans with even more branching and elongations.^{3, 7} In a typical analysis workflow, high resolution MS data helps to limit the number of possible candidate compositions, and MS/MS data are essential for determining the correct glycopeptide assignment.⁸⁻⁹ Unfortunately, collision-induced dissociation (CID), a readily available fragmentation method, has the disadvantage that it favors carbohydrate dissociation over peptide fragmentation in characterizing *O*-glycopeptides. While CID data are useful for inferring the glycan composition, typically one cannot be able to identify the peptide sequence or the glycosylation site using these data.¹⁰⁻¹¹

Among tools that can generate glycopeptide backbone cleavages including higher energy collisional dissociation (HCD)¹²⁻¹⁴, infrared multiphoton dissociation (IRMPD)^{9, 15-18} and electron capture dissociation (ECD)¹⁷⁻²⁰, electron transfer dissociation (ETD)²¹⁻²⁵ is the most widely used one and is highly orthogonal to CID in analyzing glycopeptides. Several groups have employed

ETD in large-scale glycopeptide sequencing and *O*-glycosylation site identification at the proteome level.²⁶⁻²⁸

While a number of applications are now published using ETD in glycoproteomics, automated analysis for *O*-glycopeptide data is virtually nonexistent. The only reported work for automated assignment of *O*-linked glycopeptides involved the usage of Protein Prospector software to identify glycopeptides bearing *O*-glycan structures of SA₁₋₀Hex₁₋₀HexNAc by scoring the CID and ETD spectra.^{10, 29} Nevertheless, the algorithm was developed to weight different fragment ion types based on the statistics of peptide fragmentation rules rather than glycopeptide fragmentation.³⁰⁻³¹ Moreover, only *O*-glycopeptides with simple glycan compositions could be searched and assigned.

In view of the urgent need to speed data interpretation in glycopeptide analysis, we report characteristic fragmentation patterns of intact *O*-glycopeptides in ETD and provide an algorithm to score ETD spectra of *O*-linked glycopeptides. Specifically, we show that the dominant fragment ion type in the glycopeptide sequence may vary with different precursor ions. In addition, for *O*-glycopeptide species at 3+ or higher charge state, doubly charged c^{2+} - and z^{2+} -ion series were frequently recorded in the high m/z half of the spectrum, while their singly charged counterparts were beyond the spectral mass range. These key features were incorporated into the design of an algorithm that is specifically optimized for scoring potential *O*-glycopeptide candidates against the ETD data. Using our method, site-specific assignment of *O*-glycopeptide compositions could be made in a highly accurate way. The algorithm presented here may be readily incorporated into other database search engines and data analysis tools for analyzing ETD-MS/MS spectra of *O*-glycopeptides.

3.2 Experimental Procedures

3.2.1 Samples and Reagents

Bovine fetuin was purchased from Sigma Aldrich (St. Louis, MO). The HIV envelope glycoprotein, 1086.C gp120, was expressed in transiently transfected 293T cells and purified by the Duke Human Vaccine Research Institute (Durham, NC).³² O-linked glycopeptides from erythropoietin and mucin-5AC were obtained from Anaspec (Fremont, CA) for direct-infusion MS experiments. Glycerol-free peptidyl-N-glycosidase F (PNGase F) cloned from *Flavobacterium meningosepticum* was purchased from New England BioLabs (Ipswich, MA). Sequencing grade trypsin was supplied by Promega (Madison, WI). Chemical reagents were of analytical purity or better.

3.2.2 Sample Preparation

Glycoprotein samples of 100 µg were prepared in 100 mM Tris buffer at pH 8. To remove N-glycans from the glycoprotein, samples were incubated with 2 µL of PNGase F (5000 units/mL) solution at 37 °C overnight. Subsequently, glycoproteins were denatured by 6 M urea and were treated with 5 mM tris(2-carboxyethyl)-phosphine (TCEP) to reduce the disulfide bonds. Following reduction, samples were alkylated with 20 mM iodoacetamide (IAM) at room temperature for 1 h in the dark. Excess IAM was quenched by adding 10 mM dithiothreitol (DTT). Before digestion, samples were diluted to decrease the urea concentration to 1 M. Trypsin was then added to samples at a 1:30 enzyme-to-protein ratio and the digestion lasted for 18 h at 37 °C. One microliter of formic acid was added to terminate the reaction and samples were stored at -20 °C until further analysis.

3.2.3 LC-MS/MS

Digested glycoprotein samples were subjected to online LC-MS/MS experiments. Sample was injected onto a Vydac Capillary C₈ column (320 μ m i.d. \times 10 cm, 300 Å, Micro-Tech, Vista, CA) coupled to a Thermo Scientific LTQ Velos ion trap mass spectrometer (San Jose, CA) through a Waters Acquity UltraPerformance UPLC system (Milford, MA). Mobile phases consisted of solvent A: 99.9% H₂O + 0.1% formic acid and solvent B: 99.9% CH₃CN + 0.1% formic acid. The flow rate was set at 7 μ L/min. A separation gradient was employed as follows: 5% solvent B for 5 min, followed by a linear increase to 40% B in 45 min, and then a ramp to 95% B in 10 min. The column was held at 95% B for 10 min and finally re-equilibrated in 5% B for another 15 min. For MS settings, the ESI source had a source voltage of 2.8 kV and the capillary temperature was 250 °C. Data were obtained in the positive ion mode. One sample was analyzed in two runs that were set for CID and ETD experiments, respectively. Following a full MS scan (m/z 500-2000) in the enhanced scan mode, five most intense ions from the survey scan were sequentially isolated and fragmented by CID or ETD in a data-dependent fashion. The normalized collision energy was set at 35% for CID with activation time of 10 ms. The ion-ion reaction time was 90-150 ms for ETD, and supplemental activation was turned on.³³ The automatic gain control (AGC) target value was set at 2×10^4 for the MS/MS experiment in the linear ion trap, and the AGC target value was 2×10^5 for the fluoranthene reagent anions.

3.2.4 Direct-infusion MS/MS

Glycopeptide standards having sequences of GTTPSPVPTTSTTSAP, GTTPSPVPTTSTTSAP and EAISPPDAASAAPLR (where *T* and *S* are residues modified with *N*-acetylgalactosamine, GalNAc), respectively, were dissolved in water/methanol (50:50) with 1% formic acid to a concentration of 500 nM. The prepared solution was introduced into the mass

spectrometer by direct infusion at a flow rate of 3 $\mu\text{L}/\text{min}$ in the positive ion mode. The ESI source was optimized using the following conditions: the spray voltage was 3.0 kV, capillary temperature was 200 °C and nitrogen carrier gas was 10 psi. Selected precursor ions in the full MS scan were subjected to both CID and ETD with a 2.5 Da isolation width. The activation time was 30 ms and activation energy was 30% in CID, while the reaction time in ETD was 100 ms with the maximum injection time of fluoranthene anions set as 150 ms. Thirty scans, each with 10 microscans, were averaged in the collection of MS/MS data.

3.2.5 Data Analysis

Glycoproteins used in this study (fetuin and HIV Env glycoprotein) have been well characterized in the literature regarding their *O*-glycosylation profiles and *O*-glycopeptides with known structures were searched for in the MS and MS/MS data.^{32, 34-35} Specifically, glycoproteins were tryptically digested *in silico* to produce peptides with up to 2 missed cleavages. Cysteine residues were carbamidomethylated. Theoretical masses of potential *O*-glycopeptides were calculated by adding site-specific *O*-glycan masses to the corresponding peptide sequences that contain the reported glycosylation sites. Glycopeptide masses were then converted to theoretical *m/z* values, which were searched against the full scan mass spectra with a mass tolerance of 200 ppm. If a peak was matched, the CID-MS/MS spectrum was interrogated to confirm the presence of oxonium ions [*m/z* 204 (HexNAc), 292 (SA), 366 (Hex-HexNAc), etc.] and characteristic peaks derived from monosaccharide losses. If the CID data was confirmed to be from a glycopeptide, the ETD-MS/MS spectrum of the same *O*-glycopeptide was verified manually to analyze its fragmentation patterns in ETD.

3.2.6 Algorithm Performance Test

An in-house program, GlycoPep Scorer, was coded in MATLAB based on the algorithm that was described below and in Supporting Information. A peak list file was created from each glycopeptide ETD-MS/MS spectrum and was uploaded to both GlycoPep Scorer and Protein Prospector (<http://prospector.ucsf.edu>) for scoring.³⁶ The m/z value and charge state of the precursor ion were input into each program. The glycoprotein sequence and randomized decoy sequences were directly entered into Protein Prospector. The same glycopeptide candidates were also scored by GlycoPep Scorer. Search parameters were set the same for the two programs. Trypsin was selected as the enzyme and 2 missed cleavage sites were set as the maximum. Carbamidomethylation was a fixed modification of cysteines. Mass accuracy was set to 20 ppm for precursor ions and 1.0 Da for fragment ions. In GlycoPep Scorer, the *O*-glycan composition (in the form of [SA]_n[Hex]_n[HexNAc]_n) and glycosylation site were entered for each candidate. In Protein Prospector, *O*-glycans were set as variable modifications on Ser and Thr residues in the form of HexNAc, Hex₁HexNAc, SA₁HexNAc, SA₁Hex₁HexNAc or SA₂Hex₁HexNAc. All glycopeptide identifications were manually inspected to determine the false-discovery rates.

3.3 Results and Discussion

Fetuin and the HIV Env glycoprotein, 1086.C gp120, are both *N*- and *O*-glycosylated.^{32,}
³⁴⁻³⁵ As a result, PNGase F was used to cleave the *N*-glycans off the proteins prior to tryptic digestion. In this way *N*-linked glycopeptides would not interfere with the analysis of *O*-linked glycopeptide data.²⁰ For each *O*-glycopeptide studied, its sequence, glycan composition and attachment site were confirmed based on prior knowledge of the protein, combined with the MS and CID/ETD-MS/MS data. The ETD data were specifically studied to discover distinct fragmentation patterns and to develop rules that can aid in the identification of *O*-glycopeptides.

3.3.1 *O*-Glycopeptide Fragmentation Rules in ETD

O-glycopeptide ions with m/z values over 1200 generally did not produce enough peptide fragments that could be used for sequencing. Below this value, c- and z-ions were frequently recorded in ETD spectra for glycopeptides of 2+ and higher charge states, along with y-ions and occasional peaks from glycan dissociations. However, the dominant fragment ion series varied significantly for different precursor ions, and even *O*-glycopeptides with analogous structures had distinct dissociation patterns. Figure 1A and 1B show the ETD-MS/MS data of two isomeric glycopeptides from mucin that only differ in their *O*-glycosylation sites. For the glycopeptide whose glycan attaches to Thr-3, c-ion series (c₈-c₁₄) are predominantly present in its ETD spectrum while no z-ions are found (Figure 1A). This pattern contrasts with the data from the Thr-13 glycosylated isomer, which generated both c- and z-ions during ETD (Figure 1B). Even for the same glycopeptide species, the ETD fragmentation may be drastically different if the charge state changes. The ETD spectra of an erythropoietin *O*-glycopeptide with 2+ and 3+ charges are demonstrated in Figure 1C and 1D, respectively. The doubly charged precursor ion primarily dissociates into eight dominant z-ions with only one single c-ion (c₁₄) produced in the spectrum (Figure 1C). As the glycopeptide carries more charges, however, its fragmentation efficiency in ETD improves so that both c- and z-ion series of high sequence coverage are recorded (Figure 1D). An effective algorithm for scoring *O*-glycopeptide ETD data must be optimized to score these types of spectra, where the fragment ion series is varied and unpredictable. Therefore, fixed weightings for different ion types would most likely not work optimally.

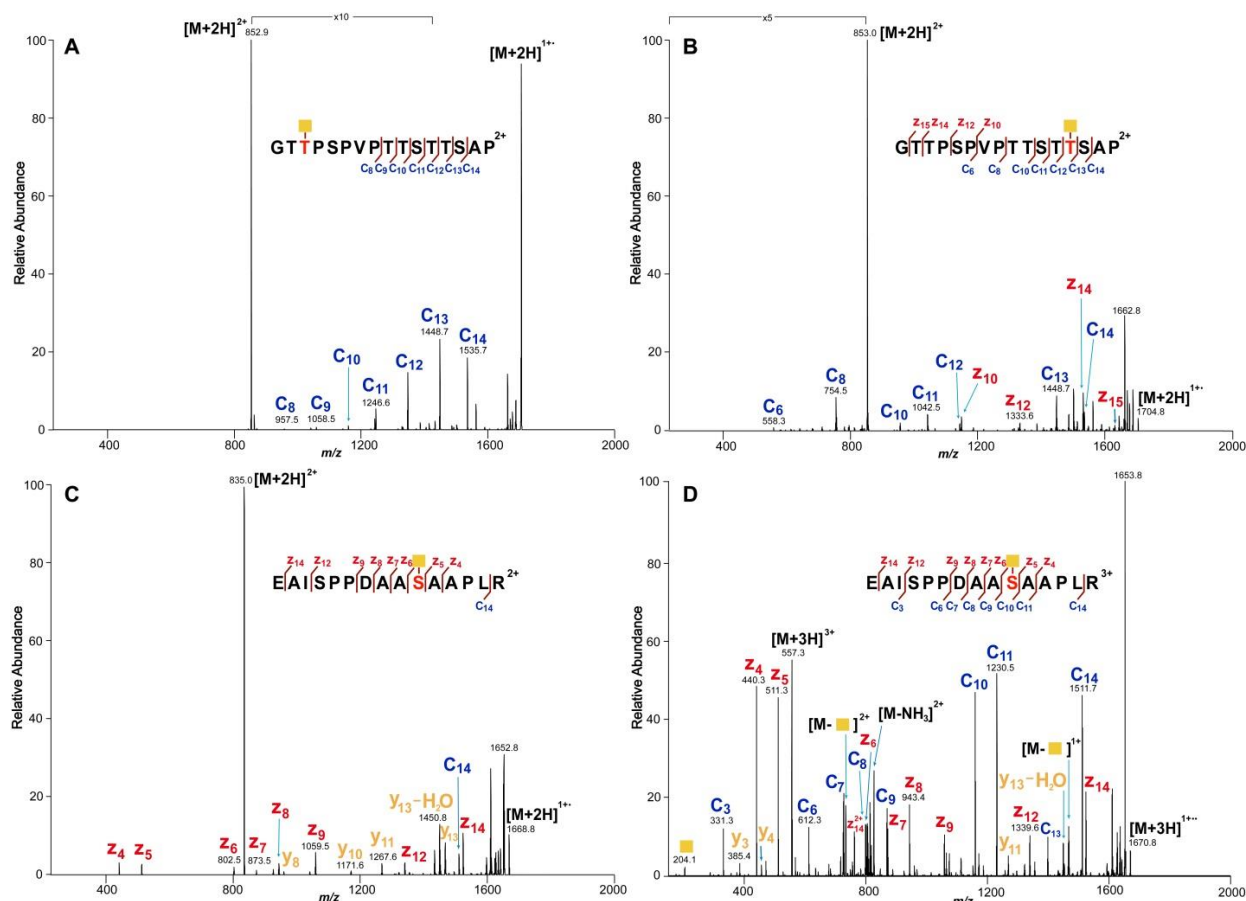


Figure 1. ETD-MS/MS data from (A) a mucin *O*-linked glycopeptide of which the glycan is attached to the Thr-3 residue (2+, m/z 852.9); (B) an isomeric *O*-glycopeptide of (A) that has the same composition but with a different modification site at Thr-13 (2+, m/z 852.9); (C) a doubly charged *O*-glycopeptide from erythropoietin with the Ser-10 residue glycosylated (m/z 834.9); (D) the same glycopeptide as (C) but at 3+ charge state (m/z 557.0). Peptide backbone fragment ions (c-, z- and y-ions) are labeled in different colors as shown in the figure. Glycan symbols used herein and in the following figures include *N*-acetylhexosamine (yellow and blue squares, HexNAc), hexose (yellow and green circles, Hex), and sialic acid (purple diamond, SA).

Furthermore, we discovered that for *O*-glycopeptides at 3+ or higher charge state, doubly charged fragment ions were likely to appear in the high m/z end of the ETD spectra. Example data are shown in Figure 2A and 2B, in which the precursor ions are two glycopeptides from the HIV envelope glycoprotein, 1086.C gp120, with 3+ and 4+ charges. For the glycopeptide in Figure 2A, the relatively large *O*-glycan modification (+1312.5 Da) makes the c-ions (c₁₂-c₁₄) and z-ions (z₇-z₁₄) that contain the glycosylated Thr-12 too large to be detected in the scan range of up to m/z

2000. Consequently, searching for these singly charged fragment ions is not very useful for increasing the coverage of the glycopeptide sequence, especially for z-ion series among which only two ions (z_3 and z_6) are found in the spectrum. By considering doubly charged c^{2+} - and z^{2+} -ions whose singly charged counterpart ions are beyond the mass range, the coverage of both c- and z-ion series are increased, as two more c^{2+} -ions and five z^{2+} -ions are identified as shown in Figure 2A. The same trend is observed in Figure 2B, where singly charged c-ions beyond c_{11} and z-ions beyond z_{10} are not recorded due to their high m/z values. However, the extra five c^{2+} -ions (c_{12}^{2+} - c_{16}^{2+}) and z^{2+} -ions (z_{10}^{2+} and z_{13}^{2+} - z_{16}^{2+}) provide extended sequence coverage for the *O*-glycopeptide. The fragmentation of *O*-glycopeptides in ETD also differs from *N*-linked glycopeptides significantly. As is illustrated in the ETD spectrum of a complex-type *N*-glycopeptide from avidin (Figure 2C), only singly charged c- and z-ion series exist while no c^{2+} - or z^{2+} -ions are generated except a single z_{16}^{2+} ion. In this circumstance, incorporating doubly charged fragment ions into the search of c- and z-ions is not helpful for identifying the correct glycopeptide composition, since it can lower the percentage of matched fragment ions over the number of possible ions being searched, and the false positive identifications would be increased.

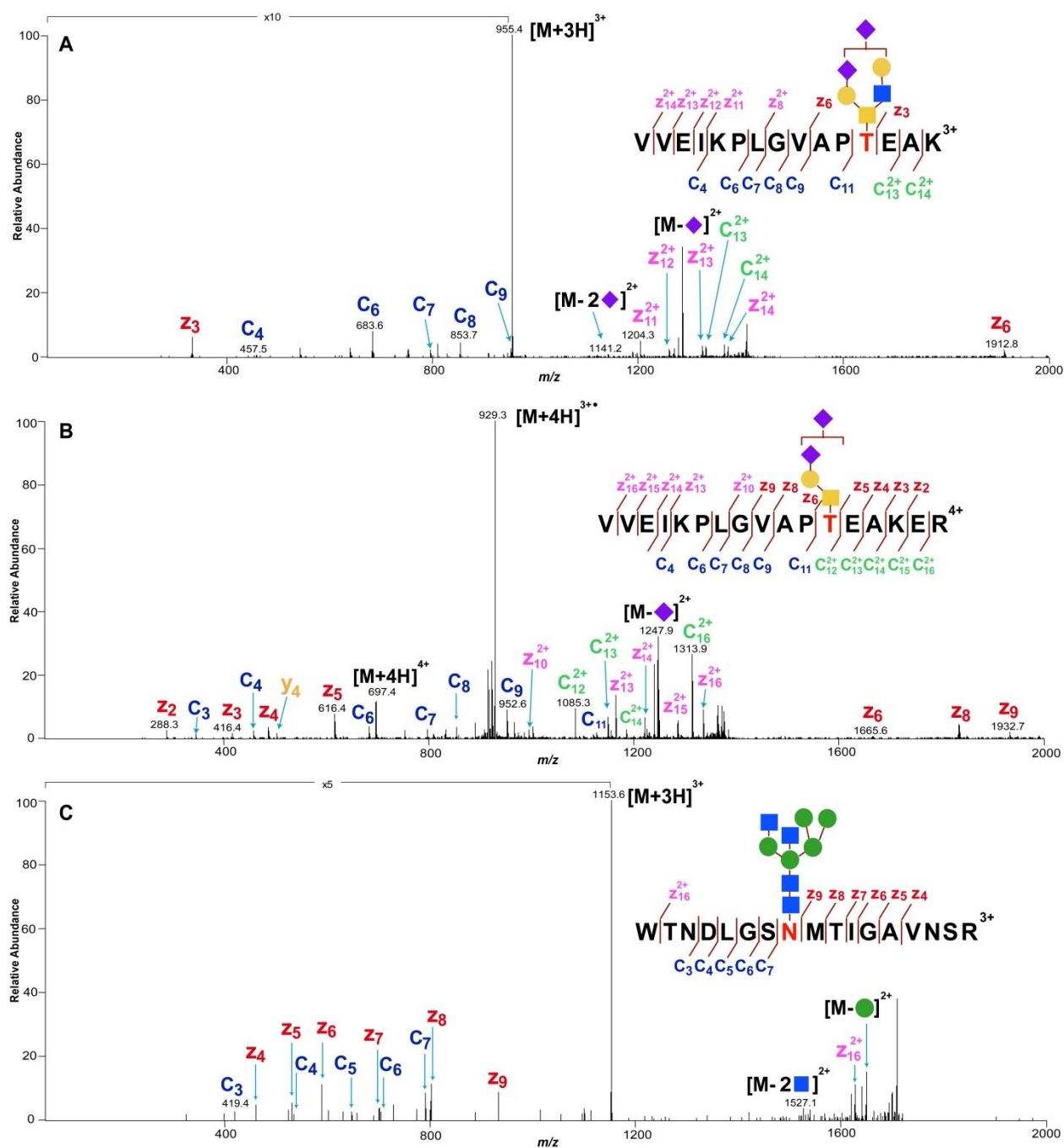


Figure 2. ETD spectra of (A) an *O*-linked core-2-type glycopeptide (3+, *m/z* 955.1) and (B) a core-1-type *O*-glycopeptide (4+, *m/z* 696.6) from the HIV envelope glycoprotein, and (C) a hybrid-type *N*-linked glycopeptide (3+, 1153.5) from avidin.

An ETD spectrum of a mucin-type core-1 *O*-glycopeptide is present in Figure 3A. The most significant spectral feature is that the major peaks in the ETD spectrum are unreacted

precursor ion, charge-reduced species and their neutral losses, which are not useful for identifying the glycopeptide sequence. However, the peptide backbone fragment ions (c- and z-ions) do exist in the data, as is illustrated in the two enlarged windows in Figure 3A, even though their relative intensities are very low compared to the base peak. Moreover, by comparing the two insets (m/z 300-500 v.s. m/z 1150-1350) in Figure 3A, it is found that interfering peaks are not evenly populated along the m/z scale in the ETD spectrum. The low m/z area has fewer peaks of noise even though the spectral intensity is low (normalized level of 2.35×10^2), while abundant interfering peaks are present in the high m/z end with relatively high intensity values (normalized level of 3.49×10^3). The trend is similar to what has been observed for *N*-linked glycopeptides, where useful fragment ions need to be differentiated from ETD side products and noise peaks.³⁷

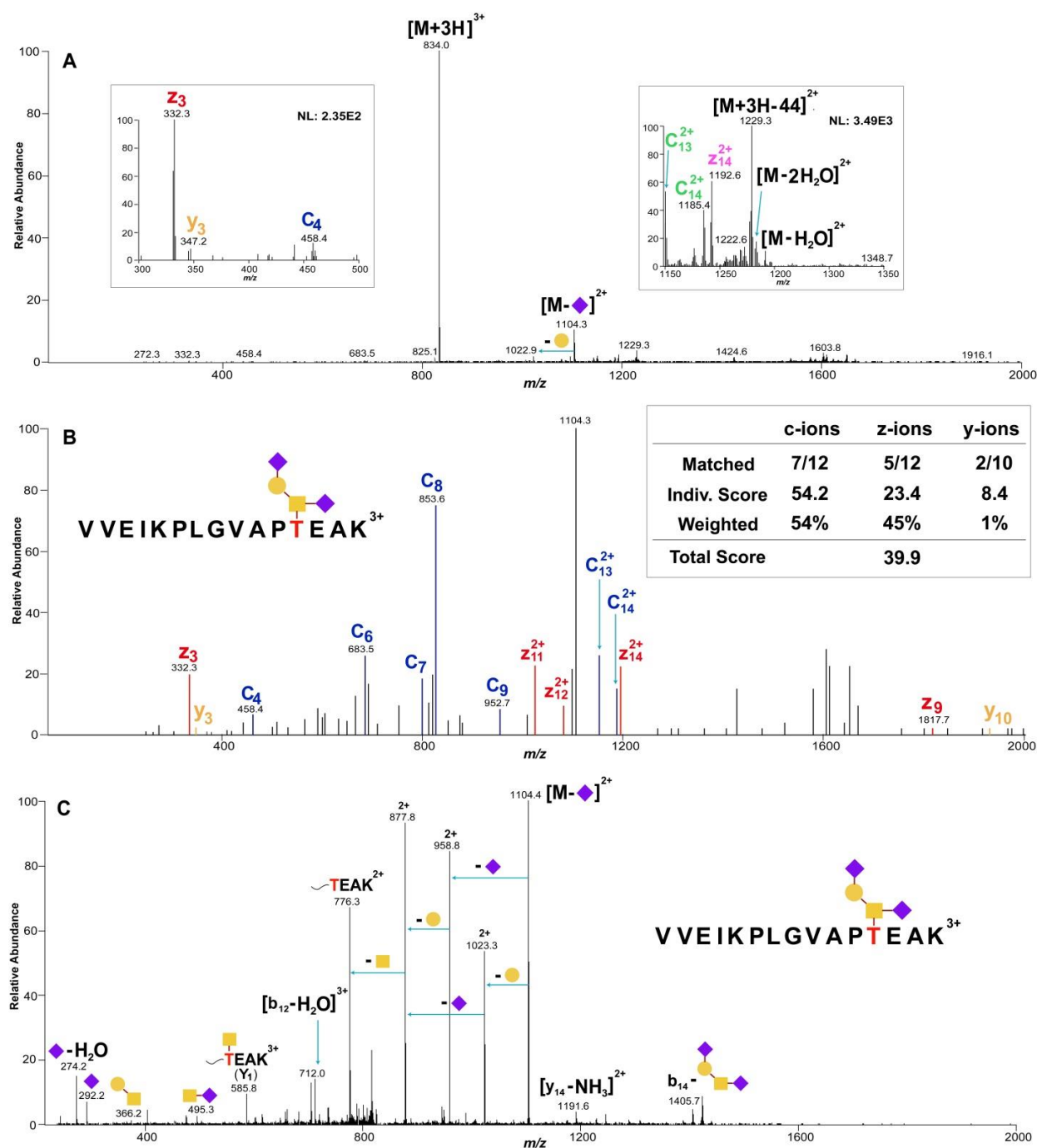


Figure 3. (A) ETD-MS/MS data of an *O*-linked glycopeptide (3+, m/z 833.4) with its composition shown in (B), note that two enlarged windows showing the zoomed m/z regions of 300-500 and of 1150-1350, respectively are also present in the figure; (B) processed ETD data of (A) after spectral filtering to remove noise peaks, and the inset table lists the scoring results (including the individual ion series' scores and the respective weightings) of the correct glycopeptide composition against the processed data; (C) CID-MS/MS data of the same glycopeptide as shown in (B).

3.3.2 Algorithm Design and Implementation in *O*-glycopeptide ETD Data Analysis

After the key features of *O*-glycopeptide fragmentation in ETD were identified, an algorithm was developed based on the characteristic fragmentation rules. First, we employed the spectral preprocessing approach that was previously used for handling *N*-glycopeptide data to filter noise peaks in the *O*-linked glycopeptide ETD spectra.³⁷ Briefly, the precursor ion, charge-reduced species and their neutral losses are removed. Subsequently, the spectrum is split into two halves by the precursor m/z value: For the low m/z half, the 5 highest peaks in every 100 Da bin are retained with other peaks removed; for the high m/z half, only the top 3 peaks are preserved in each bin. Finally, the retained peaks in the low m/z area are amplified by a factor of 5. By this method, the fragment ion peaks of low m/z values and low intensity (but good signal-to-noise ratio), as opposed to high noise peaks in the high m/z area, can be preserved and given more weighting in the scoring process.

After spectral filtering, the spectrum is then subjected to algorithm scoring. As is discussed earlier, for *O*-glycopeptides, different types of peptide backbone fragment ions have large deviations of sequence coverage in ETD, and one ion series that dominates a spectrum may not be well represented in the other spectrum. As a result, in our designed algorithm, different fragment ion series (c-, z- and y-ions) are separately searched and scored. In addition, for *O*-glycopeptides at 3+ or higher charge state, doubly charged c²⁺- and z²⁺-ions, of which the equivalent singly charged ions are beyond the scan range, are also incorporated into the search of c- and z-ion series, thus taking advantage of the distinct *O*-glycopeptide fragmentation pattern in ETD that doubly charged ions are frequently found in the high m/z end of a spectrum. It should be noted that although programs used for analyzing peptide MS/MS data also consider doubly charged fragment ions,³⁸⁻⁴⁰ our algorithm differs in that no doubly charged ion present in the low m/z half of the spectrum is searched, because these ions are typically not seen in the *O*-glycopeptide ETD data. Therefore, an individual score corresponding to each ion type, is determined by the probability that a random

sequence would have the same or higher number of matched peaks as the input glycopeptide candidate, using the following equation:

$$\text{Score}(k - \text{ion}) = -10 \times \log\left[\sum_{k=n}^N \binom{N}{k} p^k (1-p)^{N-k}\right]$$

Herein, N is the total number of searched k -ions, and n is the number of matched k -ions to the spectrum. In the next step, the weighting of each ion series is calculated by dividing the intensities of spectral peaks matched to the specific ion type into the total intensities of all matched peaks, and the total score of the candidate is then determined by summing up the weighted individual scores of c-, z- and y-ion series:

$$\text{Total Score} = \sum_{k=c, z, y} \left[\frac{\sum \text{Int. } (k - \text{ions})}{\sum \text{Int. } (all \text{ ions})} \times \text{Score}(k - \text{ion}) \right]$$

A detailed description of the algorithm, including the spectral filtering and scoring method, is contained in Supporting Information.

3.3.3 Algorithm Scoring of *O*-glycopeptide Candidate Compositions

As an example of using the algorithm in *O*-glycopeptide data analysis, the raw ETD spectrum, as illustrated in Figure 3A, was scored against the correct glycopeptide candidate of VVEIKPLGVAP**T**EAK (where **T** is modified with SA₂Hex₁HexNAc). After spectral filtering, the processed spectrum is shown in Figure 3B, where peaks that are matched to predicted fragment ions are labeled in color. The scoring parameters of all the ion series (c-, z- and y-ions) are also listed in the inset of Figure 3B. For the correct candidate, 9 singly charged c-ions and 3 doubly charged c²⁺-ions (starting from c₁₂²⁺ of m/z 1085) are searched against the processed spectrum, and 7 out of the 12 c-ions searched are matched to the ETD data. Consequently, for c-ion series, the probability that a random glycopeptide sequence has seven or more c-ions matched in the

spectrum, is calculated using the binomial distribution. An individual score of 54.2 for c-ions is then determined by converting the probability into the Log10 scale and multiplying by -10. Scores of z-ions and y-ions are computed in the same way, except that for y-ions, no doubly charged species are considered because they are not consistently produced. Subsequently, each ion series is weighted to calculate the total score of the input candidate, and the weighting factor is proportional to the matched spectral peaks' intensities, as described in detail in the algorithm in Supporting Information. As is shown in Figure 3B, multiple peaks assigned to c- and z-ions are dominant peaks in the processed spectrum, and large weightings of 54% and 45%, respectively, are given to these two ion types automatically. In contrast, the y-ion series is only 1% weighted, since only two y-ions are matched to spectral peaks of low intensity. A total score of 39.9 is then determined by summing up the product of the individual ion series' score and the weighting. Clearly, we designed the algorithm with these novel weighting features because the weighting for respective ion series varies according to the assigned spectral peaks, which is ideal for the *O*-linked glycopeptide ETD data shown here. Even when one type of ions is seriously underrepresented, that ion series' score will not have a high impact in the total score because its overall intensity level is low.

The correct glycopeptide composition, as illustrated in Figure 3B, was further corroborated by the corresponding CID spectrum, which is shown in Figure 3C. The dominant peaks in the CID data are intact peptides with sequential losses of monosaccharide units, and the glycan portion can be deduced to be SA₂Hex₁HexNAc based on these fragment ions (Figure 3C).

To test whether the algorithm is effective in differentiating the correct *O*-glycopeptide composition from multiple decoy candidates, the ETD data presented in Figure 3A, was further scored against nine isobaric decoy compositions bearing identical or similar *O*-glycan portions, and the result is summarized in Table 1. One can clearly see that the correct glycopeptide composition received the highest score of 39.9, which is significantly higher than other decoys.

Among all the incorrect assignments, the glycopeptide candidate having the sequence YKVVEIKPLGVAPTEAK (where *T* is attached to SA₁Hex₁HexNAc), received the best score of 13.8. The slightly higher score for this candidate is expected, because its sequence is highly homologous to the correct glycopeptide sequence. In this case, the algorithm still works effectively to distinguish the true candidate from the incorrect composition based on the ETD data even if their sequences are very similar.

Table 1. Algorithm scoring results of the ETD-MS/MS data against 10 *O*-glycopeptide compositions^a

Candidate	Mass	<i>O</i> -linked Glycopeptide Compositions ^b	Total Score
Correct	2497.2309	VVEIKPLGVAPTEAK + SA ₂ Hex ₁ HexNAc	39.9
Decoy 1	2497.2930	YKVVEIKPLGVAPTEAK + SA ₁ Hex ₁ HexNAc	13.8
Decoy 2	2497.2013	DFAGITGAYGAVAAGASFLFAR + Hex ₁ HexNAc	10.0
Decoy 3	2497.2224	YLTAPTITSGGNPPAFSLTSDGK + HexNAc	8.1
Decoy 4	2497.2057	ATIIVHLNESVNIK + SA ₂ Hex ₁ HexNAc	6.3
Decoy 5	2497.2309	AETPAVGLPKIEVVK + SA ₂ Hex ₁ HexNAc	5.7
Decoy 6	2497.2210	LAIQFISGNPLHK + SA ₂ Hex ₁ HexNAc	3.9
Decoy 7	2497.2516	TLFWTAVFLTIIGFGR + SA ₁ Hex ₁ HexNAc	3.4
Decoy 8	2497.2356	INSLVACGENINALLIK + SA ₁ Hex ₁ HexNAc	3.1
Decoy 9	2497.2422	GDNLLPAIVGLSILR + SA ₂ Hex ₁ HexNAc	1.9

^aThe ETD spectrum is shown in Figure 3A.

^bThe *O*-glycosylation sites are labeled in red, and the monoisotopic masses of the listed glycopeptide candidates are within 20 ppm mass error.

3.3.4 Analysis of *O*-linked Glycopeptide ETD Data Sets by GlycoPep Scorer

We integrated the spectral preprocessing method and the designed scoring algorithm into a standalone program, GlycoPep Scorer, and used the software to analyze the collected *O*-glycopeptide ETD-MS/MS data from multiple glycoproteins including mucin, fetuin, erythropoietin and the HIV envelope protein, 1086.C gp120. More than 40 ETD spectra from 22 distinct *O*-glycopeptides were scored by the program, including 5 *O*-glycopeptide species that have more than one potential glycosylation site of Ser or Thr. For every tested ETD spectrum, site-specific assignment of the corresponding *O*-glycopeptide composition was made correctly by GlycoPep Scorer, and the real glycopeptide was assigned a score at least 1.5 times higher than other decoy candidates, including positional isomers where the same peptide sequence and glycan portion are present, but the decoys differ from the correct candidate only in the glycosylation site location. The scoring results of all glycopeptide candidates using GlycoPep Scorer are summarized in Table S1 in Supporting Information.

To compare the performance of GlycoPep Scorer with other software, the same *O*-linked glycopeptide ETD data sets were also analyzed by Protein Prospector. Since the possible glycan modification is limited to simple *O*-glycan compositions, a subset of ETD spectra collected from 16 distinct glycopeptides having glycan compositions of SA₀₋₂Hex₀₋₁HexNAc, were subjected to Protein Prospector scoring. Table 2 lists the comparison of the results from the two programs. For the 16 unique *O*-glycopeptide spectra analyzed by Protein Prospector, 3 glycopeptide compositions were incorrectly assigned. In contrast, no decoy glycopeptide composition received a higher score than the correct glycopeptide candidate in GlycoPep Scorer, both for the subset data where the *O*-glycan conforms to the composition of SA₀₋₂Hex₀₋₁HexNAc, and for the whole data set, in which the glycan has a composition of SA₀₋₂Hex₀₋₂HexNAc₁₋₂. A full list of the test results

for all glycopeptide compositions scored by GlycoPep Scorer and Protein Prospector, is provided in Table S1 and S2, respectively, in Supporting Information. The raw ETD spectra are also provided in Supporting Information. The average scores for the correct composition and the best-matched decoy candidate given by both programs are presented in Table 2. For Protein Prospector, the correct glycopeptide assignment receives an average score of 44.7, and the highest decoy score averages at 29.7. A larger score difference is observed in GlycoPep Scorer for the same subset data, in which the correct composition has a score of 55.7 while the best-matched decoy composition has a score of 18.4. Although direct comparison of the absolute score values would be inappropriate since the two software's algorithms are different, GlycoPep Scorer is demonstrated herein to be efficacious in assigning site-specific *O*-glycopeptides accurately by analyzing the ETD data. Additionally, the larger score difference between the correct and incorrect assignment in GlycoPep Scorer provides more confidence that the highest scoring candidate is the right glycopeptide composition. The superior performance of the program in turn proves the advantage of using our spectral filtering approach and a scoring algorithm designed specifically for fragmentation of *O*-linked glycopeptides. At the current stage, the glycopeptide candidates need to be input manually into GlycoPep Scorer, which is probably the key drawback to using the software in its current format. However, the algorithm for scoring *O*-linked glycopeptides could be incorporated into any other glycopeptide scoring tool which uses a more automated workflow; in doing so, the convenience of automation could be combined with the power of a highly tuned scoring system.

Table 2. Analysis summary of GlycoPep Scorer and Protein Prospector in interpreting *O*-glycopeptide ETD datasets

Program Name	False Positives	Correct Score	Best Decoy Score
Protein Prospector ^a (<i>O</i> -glycan: SA ₀₋₂ Gal ₀₋₁ GalNAc ₁)	3/16	44.7	29.7
GlycoPep Detector ^a (<i>O</i> -glycan: SA ₀₋₂ Gal ₀₋₁ GalNAc ₁)	0/16	55.7	18.4
GlycoPep Detector ^b (<i>O</i> -glycan: SA ₀₋₂ Gal ₀₋₂ GalNAc ₁₋₂)	0/22	53.0	18.3

^aFalse positives and average scores were based on the scoring of a subset of 16 distinct ETD spectra collected from *O*-glycopeptides bearing glycan compositions of SA₀₋₂Gal₀₋₁GalNAc₁.

^bFalse positives and average scores were based on the scoring of a total of 22 ETD spectra collected from *O*-glycopeptides bearing glycan compositions of SA₀₋₂Gal₀₋₂GalNAc₁₋₂.

3.4 Conclusion

We studied the fragmentation of *O*-linked glycopeptides in ETD and identified their characteristic spectral features that can be applied into data analysis automation. For *O*-glycopeptides, the dominant ion series varies with different precursor ions, and it is not uncommon to see the phenomenon that one type of ion series is much more abundant than other ion types in ETD. Furthermore, we found that doubly charged c^{2+} - and z^{2+} -ions are often recorded in the high m/z half of the spectra for highly charged glycopeptides, and these ions can be included into the search of c - and z -ion series to substitute the singly charged c - and z -ions of which the m/z values are above the mass limit. In this way the sequence coverage is increased, and the individual scores of each ion series are not undermined by a lack of doubly charged ions in the low m/z end.

By correlating the weighting for each type of ions with the intensity of matched peaks, we developed an algorithm that uses *O*-glycopeptide fragmentation patterns to score the potential glycopeptide compositions against the ETD data. The algorithm, along with a spectral filtering method, was combined into the GlycoPep Scorer program, which was used in data analysis of *O*-glycopeptide ETD-MS/MS spectra. The program was able to determine the site-specific *O*-glycopeptide composition correctly with no false positives, and the large score differences between the true and decoy candidates demonstrate the benefit of the algorithm in interpreting glycopeptide ETD data. The fragmentation rules and algorithm in this study can be widely applied into other computer programs for identifying *O*-glycopeptides and determining the modification site on a large scale.

References

- (1) Van den Steen, P., Rudd, P. M., Dwek, R. A., Opdenakker, G., *Crit. Rev. Biochem. Mol. Biol.* **1998**, *33*, 151-208.
- (2) Zauner, G., Kozak, R. P., Gardner, R. A., Fernandes, D. L., Deelder, A. M., Wuhrer, M., *Biol. Chem.* **2012**, *393*, 687-708.
- (3) Jensen, P. H., Kolarich, D., Packer, N. H., *Febs J.* **2010**, *277*, 81-94.
- (4) Tarp, M. A., Clausen, H., *Biochim. Biophys. Acta-Gen. Subj.* **2008**, *1780*, 546-563.
- (5) Tian, E., Ten Hagen, K. G., *Glycoconjugate J.* **2009**, *26*, 325-334.
- (6) Burlingame, A. L., *Curr. Opin. Biotechnol.* **1996**, *7*, 4-10.
- (7) North, S. J., Hitchen, P. G., Haslam, S. M., Dell, A., *Curr. Opin. Struct. Biol.* **2009**, *19*, 498-506.
- (8) Desaire, H., Hua, D., *Int. J. Mass Spectrom.* **2009**, *287*, 21-26.
- (9) Seipert, R. R., Dodds, E. D., Lebrilla, C. B., *J. Proteome Res.* **2009**, *8*, 493-501.
- (10) Darula, Z., Chalkley, R. J., Lynn, A., Baker, P. R., Medzihradszky, K. F., *Amino Acids.* **2011**, *41*, 321-328.
- (11) Perdivara, I., Petrovich, R., Allinquant, B., Deterding, L. J., Tomer, K. B., Przybylski, M., *J. Proteome Res.* **2009**, *8*, 631-642.
- (12) Segu, Z. M., Mechref, Y., *Rapid Commun. Mass Spectrom.* **2010**, *24*, 1217-1225.
- (13) Scott, N. E., Parker, B. L., Connolly, A. M., Paulech, J., Edwards, A. V. G., Crossett, B., Falconer, L., Kolarich, D., Djordjevic, S. P., Hojrup, P., Packer, N. H., Larsen, M. R., Cordwell, S. J., *Mol. Cell. Proteomics.* **2011**, *10*, 1-18.
- (14) Hart-Smith, G., Raftery, M. J., *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 124-140.

- (15) Fukui, K., Takahashi, K., *Anal. Chem.* **2012**, *84*, 2188-2194.
- (16) Seipert, R. R., Dodds, E. D., Clowers, B. H., Beecroft, S. M., German, J. B., Lebrilla, C. B., *Anal. Chem.* **2008**, *80*, 3684-3692.
- (17) Adamson, J. T., Hakansson, K., *J. Proteome Res.* **2006**, *5*, 493-501.
- (18) Hakansson, K., Cooper, H. J., Emmett, M. R., Costello, C. E., Marshall, A. G., Nilsson, C. L., *Anal. Chem.* **2001**, *73*, 4530-4536.
- (19) Renfrow, M. B., Mackay, C. L., Chalmers, M. J., Julian, B. A., Mestecky, J., Kilian, M., Poulsen, K., Emmett, M. R., Marshall, A. G., Novak, J., *Anal. Bioanal. Chem.* **2007**, *389*, 1397-1407.
- (20) Halim, A., Ruetschi, U., Larson, G., Nilsson, J., *J. Proteome Res.* **2013**, *12*, 573-584.
- (21) Wang, D. D., Hincapie, M., Rejtar, T., Karger, B. L., *Anal. Chem.* **2011**, *83*, 2029-2037.
- (22) Snovida, S. I., Bodnar, E. D., Viner, R., Saba, J., Perreault, H., *Carbohydr. Res.* **2010**, *345*, 792-801.
- (23) Han, H., Xia, Y., Yang, M., McLuckey, S. A., *Anal. Chem.* **2008**, *80*, 3492-3497.
- (24) Thaysen-Andersen, M., Wilkinson, B. L., Payne, R. J., Packer, N. H., *Electrophoresis.* **2011**, *32*, 3536-3545.
- (25) Alley, W. R., Mechref, Y., Novotny, M. V., *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161-170.
- (26) Chalkley, R. J., Thalhammer, A., Schoepfer, R., Burlingame, A. L., *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 8894-8899.
- (27) Steentoft, C., Vakhrushev, S. Y., Vester-Christensen, M. B., Schjoldager, K., Kong, Y., Bennett, E. P., Mandel, U., Wandall, H., Levery, S. B., Clausen, H., *Nat. Methods.* **2011**, *8*, 977-982.
- (28) Vakhrushev, S. Y., Steentoft, C., Vester-Christensen, M. B., Bennett, E. P., Clausen, H., Levery, S. B., *Mol. Cell. Proteomics.* **2013**, *12*, 932-944.
- (29) Darula, Z., Chalkley, R. J., Baker, P., Burlingame, A. L., Medzihradszky, K. F., *Eur. J. Mass Spectrom.* **2010**, *16*, 421-428.

- (30) Chalkley, R. J., Medzihradszky, K. F., Lynn, A. J., Baker, P. R., Burlingame, A. L., *Anal. Chem.* **2010**, 82, 579-584.
- (31) Baker, P. R., Medzihradszky, K. F., Chalkley, R. J., *Mol. Cell. Proteomics.* **2010**, 9, 1795-1803.
- (32) Go, E. P., Liao, H. X., Alam, S. M., Hua, D., Haynes, B. F., Desaire, H., *J. Proteome Res.* **2013**, 12, 1223-1234.
- (33) Swaney, D. L., McAlister, G. C., Wirtala, M., Schwartz, J. C., Syka, J. E. P., Coon, J. J., *Anal. Chem.* **2007**, 79, 477-485.
- (34) Carr, S. A., Huddleston, M. J., Bean, M. F., *Protein Sci.* **1993**, 2, 183-196.
- (35) Nwosu, C. C., Seipert, R. R., Strum, J. S., Hua, S. S., An, H. J., Zivkovic, A. M., German, B. J., Lebrilla, C. B., *J. Proteome Res.* **2011**, 10, 2612-2624.
- (36) Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., Burlingame, A. L., *Mol. Cell. Proteomics.* **2005**, 4, 1194-1204.
- (37) Zhu, Z., Hua, D., Clark, D. F., Go, E. P., Desaire, H., *Anal. Chem.* **2013**, 85, 5023-32.
- (38) Hogan, J. M., Higdon, R., Kolker, N., Kolker, E., *Omics.* **2005**, 9, 233-250.
- (39) Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X. Y., Shi, W. Y., Bryant, S. H., *J. Proteome Res.* **2004**, 3, 958-964.
- (40) Deutsch, E. W., Shteynberg, D., Lam, H., Sun, Z., Eng, J. K., Carapito, C., von Haller, P. D., Tasman, N., Mendoza, L., Farrah, T., Aebersold, R., *PROTEOMICS.* **2010**, 10, 1190-1195.

Table S1. Score results by GlycoPep Scorer for *O*-linked glycopeptide ETD data set.^A

Test	Charge	Candidate ^B	<i>m/z</i> ^C	Glycopeptide Composition ^D	Score ^E
1	3+	Correct	711.0266	TPIVGQPSIPGGPVR + [SA]1[Hex]1[HexNAc]1	55.4
	3+	Decoy 1	711.0266	TPIVGQPSIPGGPVR + [SA]1[Hex]1[HexNAc]1	17.5
	3+	Decoy 2	711.0266	VPGGPISPQGVIPTR + [SA]1[Hex]1[HexNAc]1	7.8
	3+	Decoy 3	711.0094	NKTDELSEKELAAAR + [SA]1[Hex]1[HexNAc]1	7.1
	3+	Decoy 4	711.0094	AALEKSELEDTKNR + [SA]1[Hex]1[HexNAc]1	6.4
2	3+	Correct	928.4655	VVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1	71.5
	3+	Decoy 1	928.4865	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	22.9
	3+	Decoy 2	928.4596	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^F	11.4
	3+	Decoy 3	928.4655	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1	6.9
	3+	Decoy 4	928.4596	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^F	3.3
3	4+	Correct	696.6009	VVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1	80.2
	4+	Decoy 1	696.6167	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	24.6
	4+	Decoy 2	696.5965	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^F	14.5
	4+	Decoy 3	696.6009	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1	13.4
	4+	Decoy 4	696.5965	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^F	2.8
4	3+	Correct	833.4176	VVEIKPLGVAPTEAK + [SA]2[Hex]1[HexNAc]1	39.9
	3+	Decoy 1	833.4386	YKVVEIKPLGVAPTEAK + [SA]1[Hex]1[HexNAc]1	13.8
	3+	Decoy 2	833.4077	DFAGITGAYGAVAAGASFLFAR + [Hex]1[HexNAc]1	10.0

	3+	Decoy 3	833.4148	YLTAPTITSGGNPPAFSLTSDGK + [HexNAc]1	8.1	
	3+	Decoy 4	833.4092	ATIIVHLNESVNIK + [SA]2[Hex]1[HexNAc]1	6.3	
	3+	Decoy 5	833.4176	AETPAVGLPKIEVVK + [SA]2[Hex]1[HexNAc]1	5.7	
	3+	Decoy 6	833.4143	LAIIQFISGNPLHK + [SA]2[Hex]1[HexNAc]1	3.9	
	3+	Decoy 7	833.4245	TLFWTAVFLTIIGFGR + [SA]1[Hex]1[HexNAc]1	3.4	
	3+	Decoy 8	833.4192	INSLVACGENINALLIK + [SA]1[Hex]1[HexNAc]1	3.1	
	3+	Decoy 9	833.4214	GDNLLPAIVGLSILR + [SA]2[Hex]1[HexNAc]1	1.9	
	5	2+	Correct	958.5273	VVEIKPLGVAPTEAK + [Hex]1[HexNAc]1	38.3
	2+	Decoy 1	958.5164	ATIIVHLNESVNIK + [Hex]1[HexNAc]1	19.9	
	2+	Decoy 2	958.5273	AETPAVGLPKIEVVK + [Hex]1[HexNAc]1	18.8	
	2+	Decoy 3	833.4214	ATIIVHLNESVNIK + [Hex]1[HexNAc]1	12.8	
	6	2+	Correct	1125.5779	VVEIKPLGVAPTEAK + [SA]1[Hex]1[HexNAc]1 ^G	23.5
	2+	Decoy 1	1125.5779	AETPAVGLPKIEVVK + [SA]1[Hex]1[HexNAc]1 ^G	4.6	
	2+	Decoy 2	1125.5585	MQFNIPTLLTLFR + [SA]1[Hex]1[HexNAc]1	4.2	
	2+	Decoy 3	1125.6094	YKVVEIKPLGVAPTEAK + [Hex]1[HexNAc]1 ^G	2.7	
	2+	Decoy 4	1125.5347	MDFLLEALTNWLK + [SA]1[Hex]1[HexNAc]1	0	
	7	2+	Correct	1246.6469	VVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	26.6
	2+	Decoy 1	1246.6380	VQCTHGIPVVSTQLLLK + [SA]1[Hex]1[HexNAc]1 ^F	15.4	
	2+	Decoy 2	1246.6783	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	11.6	
	2+	Decoy 3	1246.6469	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	6.9	
	2+	Decoy 4	1246.6632	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	6.3	
	2+	Decoy 5	1246.6380	VQCTHGIPVVSTQLLLK + [SA]1[Hex]1[HexNAc]1 ^F	0	
	8	3+	Correct	831.4337	VVEIKPLGVAPTEAKER +	72.6

				[SA]1[Hex]1[HexNAc]1	
	3+	Decoy 1	831.4547	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	16.5
	3+	Decoy 2	831.4337	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	13.5
	3+	Decoy 3	831.4278	VQCTHGIPVVSTQLLK + [SA]1[Hex]1[HexNAc]1 ^F	12.5
	3+	Decoy 4	831.4446	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	10.8
	3+	Decoy 5	831.4278	VQCTHGIPVVSTQLLK + [SA]1[Hex]1[HexNAc]1 ^F	9.5
9	4+	Correct	623.8271	VVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	76.8
	4+	Decoy 1	623.8428	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	16.8
	4+	Decoy 2	623.8271	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	12.6
	4+	Decoy 3	623.8227	VQCTHGIPVVSTQLLK + [SA]1[Hex]1[HexNAc]1 ^F	4.4
	4+	Decoy 4	623.8227	VQCTHGIPVVSTQLLK + [SA]1[Hex]1[HexNAc]1 ^F	3.0
	4+	Decoy 5	623.8353	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	2.5
10	3+	Correct	942.8008	VVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1 ^G	42.4
	3+	Decoy 1	942.8217	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1 ^G	18.8
	3+	Decoy 2	942.8008	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1 ^G	15.3
	3+	Decoy 3	942.7962	CALSVLANLLLQFAFAK + [SA]2[Hex]1[HexNAc]1	6.1
	3+	Decoy 4	942.7984	FILVSLPELVLNQPAPE + [SA]2[Hex]1[HexNAc]1	3.9
11	3+	Correct	1039.8535	YKVVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1 ^G	37.5

	3+	Decoy 1	1039.8535	EKAETPAVGLPKIEVVKYR + [SA]2[Hex]1[HexNAc]1 ^G	16.8
	3+	Decoy 2	1039.8619	EALPIHLLDGVPLLAGSVDIK + [SA]2[Hex]1[HexNAc]1	4.3
	3+	Decoy 3	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	0
	3+	Decoy 4	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	0
12	4+	Correct	780.1420	YKVVEIKPLGVAPTAKER + [SA]2[Hex]1[HexNAc]1 ^G	71.3
	4+	Decoy 1	780.1420	EKAETPAVGLPKIEVVKYR + [SA]2[Hex]1[HexNAc]1 ^G	6.8
	4+	Decoy 2	780.1482	EALPIHLLDGVPLLAGSVDIK + [SA]2[Hex]1[HexNAc]1	4.2
	4+	Decoy 3	780.1420	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	1.7
	4+	Decoy 4	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	1.2
13	2+	Correct	834.9256	EAISPPDAASAAPLR + [HexNAc]1	72.6
	2+	Decoy 1	834.9256	EAISSPPDAASAAPLR + [HexNAc]1	44.9
	2+	Decoy 2	834.9223	MTVDQQVAHAIPR + [HexNAc]1	7.7
	2+	Decoy 3	834.9167	MGDNSLASFIVGVR + [HexNAc]1	6.1
	2+	Decoy 4	834.9167	MGDNSLASFIVGVR + [HexNAc]1	3.2
	2+	Decoy 5	834.9111	EISGVFAAGDICVK + [HexNAc]1	2.2
14	2+	Decoy 6	834.9276	SAFLPIEDAYAIR + [HexNAc]1	0
	3+	Correct	556.9528	EAISPPDAASAAPLR + [HexNAc]1	65.8
	3+	Decoy 1	556.9528	EAISSPPDAASAAPLR + [HexNAc]1	27.6
	3+	Decoy 2	556.9542	SAFLPIEDAYAIR + [HexNAc]1	14.3
	3+	Decoy 3	556.9469	MGDNSLASFIVGVR + [HexNAc]1	9.9
	3+	Decoy 4	556.9469	MGDNSLASFIVGVR + [HexNAc]1	7.9

15	3+	Decoy 5	556.9506	MTVDQQVAHAIPR + [HexNAc]1	4.5
	3+	Decoy 6	556.9432	EISGVFAAGDICVK + [HexNAc]1	3.5
	2+	Correct	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	77.2
	2+	Decoy 1	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	21.7
	2+	Decoy 2	852.9005	WVCELFYDTLR + [HexNAc]1	9.0
	2+	Decoy 3	852.9074	FHLGTTELDPTDR + [HexNAc]1	4.2
	2+	Decoy 4	852.9005	QMYIVYNTIGYH + [HexNAc]1	4.2
16	2+	Decoy 5	852.9131	MNFSELIPFFEK + [HexNAc]1	0
	2+	Correct	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	46.7
	2+	Decoy 1	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	8.4
	2+	Decoy 2	852.9005	QMYIVYNTIGYH + [HexNAc]1	3.1
	2+	Decoy 3	852.9131	MNFSELIPFFEK + [HexNAc]1	3.1
	2+	Decoy 4	852.9074	FHLGTTELDPTDR + [HexNAc]1	2.8
	2+	Decoy 5	852.9005	WVCELFYDTLR + [HexNAc]1	2.2
17	3+	Correct	955.1283	VVEIKPLGVAPTEAK + [SA]2[Hex]2[HexNAc]2	48.2
	3+	Decoy 1	955.1283	AETPAVGLPKIEVVK + [SA]2[Hex]2[HexNAc]2	8.1
	3+	Decoy 2	955.1199	ATIIVHLNESVNIK + [SA]2[Hex]2[HexNAc]2	6.7
	3+	Decoy 3	955.1199	ATIIVHLNESVNIK + [SA]2[Hex]2[HexNAc]2	6.7
	3+	Decoy 4	955.1493	YKVVEIKPLGVAPTEAK + [SA]1[Hex]2[HexNAc]2	6.4
18	4+	Correct	1050.1762	VVEIKPLGVAPTEAKER + [SA]2[Hex]2[HexNAc]2	36.8
	4+	Decoy 1	1050.1972	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]2[HexNAc]2	15.8
	4+	Decoy 2	1050.1703	VQCTHGIPVVSTQLLLK + [SA]2[Hex]2[HexNAc]2 ^F	9.6
	4+	Decoy 3	1050.1703	VQCTHGIPVVSTQLLLK + [SA]2[Hex]2[HexNAc]2 ^F	2.5
	4+	Decoy 4	1050.1762	EKAETPAVGLKPIEVVR + [SA]2[Hex]2[HexNAc]2	2.0

19	3+	Correct	953.1444	VVEIKPLGVAPTEAKER + [SA]1[Hex]2[HexNAc]2	49.8
	3+	Decoy 1	953.1654	YKVVEIKPLGVAPTEAKER + [Hex]2[HexNAc]2	31.5
	3+	Decoy 2	953.1385	VQCTHGIPVVSTQLLLK + [SA]1[Hex]2[HexNAc]2 ^F	8.8
	3+	Decoy 3	953.1385	VQCTHGIPVVSTQLLLK + [SA]1[Hex]2[HexNAc]2 ^F	7.3
	3+	Decoy 4	953.1444	EKAETPAVGLKPIEVVR + [SA]1[Hex]2[HexNAc]2	5.7
20	4+	Correct	715.1101	VVEIKPLGVAPTEAKER + [SA]1[Hex]2[HexNAc]2	56.2
	4+	Decoy 1	715.1529	YKVVEIKPLGVAPTEAKER + [Hex]2[HexNAc]2	28.4
	4+	Decoy 2	715.1101	EKAETPAVGLKPIEVVR + [SA]1[Hex]2[HexNAc]2	16.7
	4+	Decoy 3	715.1057	VQCTHGIPVVSTQLLLK + [SA]1[Hex]2[HexNAc]2 ^F	4.3
	4+	Decoy 4	715.1057	VQCTHGIPVVSTQLLLK + [SA]1[Hex]2[HexNAc]2 ^F	2.5
21	3+	Correct	967.4797	VVEIKPLGVAPTEAKER + [SA]1[Hex]2[HexNAc]2 ^G	50.5
	3+	Decoy 1	967.5007	YKVVEIKPLGVAPTEAKER + [Hex]2[HexNAc]2 ^G	14.6
	3+	Decoy 2	967.4797	EKAETPAVGLKPIEVVR + [SA]1[Hex]2[HexNAc]2 ^G	10.4
	3+	Decoy 3	967.4773	FILVSLPELVLNQPAPE + [SA]1[Hex]2[HexNAc]2	3.5
	3+	Decoy 4	967.4751	CALSVLANLLLQFAFAK + [SA]1[Hex]2[HexNAc]2	0
22	3+	Correct	1161.5643	YKVVEIKPLGVAPTEAKER + [SA]2[Hex]2[HexNAc]2 ^G	32.5
	3+	Decoy 1	1161.5726	EALPIHLLDGVPLLAGSVDIK + [SA]2[Hex]2[HexNAc]2	9.5
	3+	Decoy 2	1161.5643	EKAETPAVGLKPIEVVR + [SA]2[Hex]2[HexNAc]2 ^G	7.2

3+	Decoy 3	1161.5681	MTAAFIG T LLALIVYNAFLK + [SA]2[Hex]2[HexNAc]2	1.6
----	---------	-----------	---	-----

^A For Test 1-16, the correct glycopeptide composition has the *O*-glycan portion of [SA]₀₋₂[Hex]₀₋₁[HexNAc]₁. For Test 17-22, the correct candidate has the glycan of [SA]₀₋₂[Hex]₀₋₂[HexNAc]₁₋₂. The raw spectra of Test 1-22 are summarized in an Excel file in Supporting Information.

^B For the scored candidates in each test, the correct candidate is always listed at the top, while decoy candidates are listed by their total scores in a high to low order.

^C The monoisotopic *m/z* values of the scored glycopeptide candidates are within 20 ppm mass error from the MS¹ *m/z* values.

^D The *O*-glycosylation site is labeled in red for each input glycopeptide candidate.

^E The score shown in the table is the total score for each glycopeptide composition calculated based on the algorithm described in the manuscript.

^F The cysteine residue is not alkylated in the corresponding glycopeptide composition.

^G Carbamylation occurs on the N-terminus of the corresponding glycopeptide composition.

Table S2. Score results by Protein Prospector for O-glycopeptide ETD data set.^A

Test	Charge	Candidate ^B	<i>m/z</i>	Glycopeptide Composition	Score ^C
1	3+	Correct	711.0266	TPIVGQPS I PGGPVR + [SA]1[Hex]1[HexNAc]1	53.8
	3+	Decoy 1	711.0266	T PIVGQPSIPGGPVR + [SA]1[Hex]1[HexNAc]1	29.0
	3+	Decoy 2	711.0266	VPGGPI S PQGVIPT R + [SA]1[Hex]1[HexNAc]1	-3.6
	3+	Decoy 3	711.0094	NKTDEL S KELAAR + [SA]1[Hex]1[HexNAc]1	4.8
	3+	Decoy 4	711.0094	AALEK S LEDTKNR + [SA]1[Hex]1[HexNAc]1	-4.8
2	3+	Correct	928.4655	VVEIKPLGVAP T EAKER + [SA]2[Hex]1[HexNAc]1	24.1
	3+	Decoy 1	928.4865	YKVVEIKPLGVAP T EAKER + [SA]1[Hex]1[HexNAc]1	13.0

3	3+	Decoy 2	928.4596	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^D	-10.8
	3+	Decoy 3	928.4655	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1	-4.7
	3+	Decoy 4	928.4596	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^D	-14.1
	4+	Correct	696.6009	VVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1	41.5
	4+	Decoy 1	696.6167	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	33.3
4	4+	Decoy 2	696.5965	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^D	1.6
	4+	Decoy 3	696.6009	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1	29.8
	4+	Decoy 4	696.5965	VQCTHGIPVVSTQLLLK + [SA]2[Hex]1[HexNAc]1 ^D	16.4
	3+	Correct	833.4176	VVEIKPLGVAPTEAK + [SA]2[Hex]1[HexNAc]1	4.5
	3+	Decoy 1	833.4386	YKVVEIKPLGVAPTEAK + [SA]1[Hex]1[HexNAc]1	17.1
5	3+	Decoy 2	833.4077	DFAGITGAYGAVAAGASFLFAR + [Hex]1[HexNAc]1	28.3
	3+	Decoy 3	833.4148	YLTAPTITSGGNPPAFSLTSDGK + [HexNAc]1	31.7
	3+	Decoy 4	833.4092	ATIIVHLNESVNIK + [SA]2[Hex]1[HexNAc]1	-28.8
	3+	Decoy 5	833.4176	AETPAVGLPKIEVVK + [SA]2[Hex]1[HexNAc]1	-24.0
	3+	Decoy 6	833.4143	LAIIQFISGNPLHK + [SA]2[Hex]1[HexNAc]1	-69.5
	3+	Decoy 7	833.4245	TLFWTAVFLTIIIGFGR + [SA]1[Hex]1[HexNAc]1	-20.7
	3+	Decoy 8	833.4192	INSLVACGENINALLIK + [SA]1[Hex]1[HexNAc]1	4.3
	3+	Decoy 9	833.4214	GDNLLPAIVGLSILR + [SA]2[Hex]1[HexNAc]1	-24.1
	2+	Correct	958.5273	VVEIKPLGVAPTEAK + [Hex]1[HexNAc]1	80.5
5	2+	Decoy 1	958.5164	ATIIVHLNESVNIK + [Hex]1[HexNAc]1	3.5
	2+	Decoy 2	958.5273	AETPAVGLPKIEVVK + [Hex]1[HexNAc]1	47.3
	2+	Decoy 3	833.4214	ATIIVHLNESVNIK + [Hex]1[HexNAc]1	23.5

6	2+	Correct	1125.5779	VVEIKPLGVAPTEAK + [SA]1[Hex]1[HexNAc]1 ^E	29.8
	2+	Decoy 1	1125.5779	AETPAVGLPKIEVVK + [SA]1[Hex]1[HexNAc]1 ^E	-13.8
	2+	Decoy 2	1125.5585	MQFNIPTLLTLFR + [SA]1[Hex]1[HexNAc]1	-0.5
	2+	Decoy 3	1125.6094	YKVVEIKPLGVAPTEAK + [Hex]1[HexNAc]1 ^E	6.1
	2+	Decoy 4	1125.5347	MDFLLEALT ^N NWLK + [SA]1[Hex]1[HexNAc]1	-15.4
7	2+	Correct	1246.6469	VVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	11.1
	2+	Decoy 1	1246.6380	VQCTHGIPVVSTQ ^L LLK + [SA]1[Hex]1[HexNAc]1 ^D	6.9
	2+	Decoy 2	1246.6783	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	3.9
	2+	Decoy 3	1246.6469	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	-4.5
	2+	Decoy 4	1246.6632	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	5.1
8	3+	Correct	831.4337	VVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	54.1
	3+	Decoy 1	831.4547	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	30.3
	3+	Decoy 2	831.4337	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	-6.1
	3+	Decoy 3	831.4278	VQCTHGIPVVSTQ ^L LLK + [SA]1[Hex]1[HexNAc]1 ^D	-4.4
	3+	Decoy 4	831.4446	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	-11.9
9	4+	Correct	623.8271	VVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1	85.0
	4+	Decoy 1	623.8428	YKVVEIKPLGVAPTEAKER + [Hex]1[HexNAc]1	53.2

	4+	Decoy 2	623.8271	EKAETPAVGLKPIEVVR + [SA]1[Hex]1[HexNAc]1	25.9
	4+	Decoy 3	623.8227	VQCTHGIPVVSTQLLLK + [SA]1[Hex]1[HexNAc]1 ^F	8.8
	4+	Decoy 4	623.8227	VQCTHGIPVVSTQLLLK + [SA]1[Hex]1[HexNAc]1 ^F	2.4
	4+	Decoy 5	623.8353	MALPLEILIVPESLIGK + [SA]1[Hex]1[HexNAc]1	-5.2
	3+	Correct	942.8008	VVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1 ^E	15.2
10	3+	Decoy 1	942.8217	YKVVEIKPLGVAPTEAKER + [SA]1[Hex]1[HexNAc]1 ^E	9.4
	3+	Decoy 2	942.8008	EKAETPAVGLKPIEVVR + [SA]2[Hex]1[HexNAc]1 ^E	-14.0
	3+	Decoy 3	942.7962	CALSVLANLLLQFAFAK + [SA]2[Hex]1[HexNAc]1	-15.6
	3+	Decoy 4	942.7984	FILVSLPELVLNQPAPE + [SA]2[Hex]1[HexNAc]1	-64.2
11	3+	Correct	1039.8535	YKVVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1 ^E	51.7
	3+	Decoy 1	1039.8535	EKAETPAVGLPKIEVVKYR + [SA]2[Hex]1[HexNAc]1 ^E	5.5
	3+	Decoy 2	1039.8619	EALPIHLLDGVPLLAGSVDIK + [SA]2[Hex]1[HexNAc]1	5.3
	3+	Decoy 3	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	-30.7
	3+	Decoy 4	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	18.1
12	4+	Correct	780.1420	YKVVEIKPLGVAPTEAKER + [SA]2[Hex]1[HexNAc]1 ^E	69.0
	4+	Decoy 1	780.1420	EKAETPAVGLPKIEVVKYR + [SA]2[Hex]1[HexNAc]1 ^E	39.0
	4+	Decoy 2	780.1482	EALPIHLLDGVPLLAGSVDIK +	3.6

			[SA]2[Hex]1[HexNAc]1		
	4+	Decoy 3	780.1420	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	53.1
	4+	Decoy 4	1039.8573	MTAAFIGTLLALIVYNAFLK + [SA]2[Hex]1[HexNAc]1	-15.5
13	2+	Correct	834.9256	EAI SPDAAS A APLR + [HexNAc]1	77.2
	2+	Decoy 1	834.9256	EAI SPDAAS A APLR + [HexNAc]1	55.3
	2+	Decoy 2	834.9223	MTVDQQVAHAIPR + [HexNAc]1	15.1
	2+	Decoy 3	834.9167	MGDN SLASFIVGVR + [HexNAc]1	21.7
	2+	Decoy 4	834.9167	MGDN SLASFIVGVR + [HexNAc]1	7.2
	2+	Decoy 5	834.9111	EI SGVFAAGDICVK + [HexNAc]1	18.7
	2+	Decoy 6	834.9276	SAFLPIEDAYAIR + [HexNAc]1	-5.2
14	3+	Correct	556.9528	EAI SPDAAS A APLR + [HexNAc]1	71.2
	3+	Decoy 1	556.9528	EAI SPDAAS A APLR + [HexNAc]1	76.6
	3+	Decoy 2	556.9542	SAFLPIEDAYAIR + [HexNAc]1	11.0
	3+	Decoy 3	556.9469	MGDN SLASFIVGVR + [HexNAc]1	52.4
	3+	Decoy 4	556.9469	MGDN SLASFIVGVR + [HexNAc]1	63.0
	3+	Decoy 5	556.9506	MTVDQQVAHAIPR + [HexNAc]1	27.0
	3+	Decoy 6	556.9432	EI SGVFAAGDICVK + [HexNAc]1	49.0
15	2+	Correct	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	57.6
	2+	Decoy 1	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	23.1
	2+	Decoy 2	852.9005	WVCELFYD TLR + [HexNAc]1	26.8
	2+	Decoy 3	852.9074	FHLGT TELDPTDR + [HexNAc]1	0.3
	2+	Decoy 4	852.9005	QMYIVYNTIGYH + [HexNAc]1	7.7
	2+	Decoy 5	852.9131	MNF SELIPFEK + [HexNAc]1	-14.9
16	2+	Correct	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	-11.2
	2+	Decoy 1	852.9124	GTTPSPVPTTSTTSAP + [HexNAc]1	-8.8
	2+	Decoy 2	852.9005	QMYIVYNTIGYH + [HexNAc]1	-16.7

2+	Decoy 3	852.9131	MNFS S ELIPFFEK + [HexNAc]1	-5.3
2+	Decoy 4	852.9074	FHLGT T ELDPTDR + [HexNAc]1	-3.8
2+	Decoy 5	852.9005	WVCELFYD T LR + [HexNAc]1	-14.0

^A The ETD data analyzed by Protein Prospector is the same data that has been scored by GlycoPep Scorer in Test 1-16. For this subset of data, the correct glycopeptide composition has the *O*-glycan portion of [SA]₀₋₂[Hex]₀₋₁[HexNAc]₁.

^B The same glycopeptide candidates scored by GlycoPep Scorer in Test 1-16 are analyzed by Protein Prospector herein in the same order. Note that in each test, the correct candidate is still listed at the top.

^C The score value for each candidate is given by the Protein Prospector software. Note that in three cases (Test 4, 14 and 16), at least one of the decoy candidates received a higher score than the correct candidate.

^D The cysteine residue is not alkylated in the corresponding glycopeptide composition.

^E Carbamylation occurs on the N-terminus of the corresponding glycopeptide composition.

GlycoPep Scorer Algorithm

(As of time of publication, subject to revision for enhancement.)

INPUTS:

Spectrum = Raw ETD-MS/MS data

PrecursorIon = m/z of the precursor ion

ChargeState = charge state of the precursor ion

Charge-reducedSpecies = m/z of the charge-reduced species

Spectrum[Minima] = lower scan limit

Spectrum[Maxima] = upper scan limit

Candidate *O*-glycan and Peptide Formulas (with the glycosylation site indicated)

[A] Spectral filtering process:

1A. Removal of the precursor ion peak and charge-reduced species' peaks

a. Calculate the m/z of the charge-reduced species:

Charge-reducedSpecies at ChargeState-1 = $(\text{PrecursorIon} \times \text{ChargeState}) \div (\text{ChargeState} - 1)$

Charge-reducedSpecies at ChargeState-2 = $(\text{PrecursorIon} \times \text{ChargeState}) \div (\text{ChargeState} - 2)$

.....

Charge-reducedSpecies at ChargeState- n = $(\text{PrecursorIon} \times \text{ChargeState}) \div (\text{ChargeState} - n)$

If Charge-reducedSpecies at ChargeState- n > Spectrum[Maxima], the calculation procedure is ended.

b. Removal of peaks from the precursor ion and charge-reduced species:

All peaks within PrecursorIon ± 3 Da range are deleted.

All peaks from {Charge-reducedSpecies - $[60 \div (\text{ChargeState} - n)] - 2$ } to {Charge-reducedSpecies + 3} are deleted.

2A. Peaks that are less than 3 Da apart are binned together and compared with each other in intensity. Only the peak of the highest intensity is retained while others are deleted in each bin.

3A. The spectrum is split into two halves by PrecursorIon:

For the first half from Spectrum[Minima] to PrecursorIon, in every 100 Da window, the top five peaks in intensity are retained while other peaks are deleted.

For the second half from PrecursorIon to Spectrum[Maxima], in every 100 Da window, the top three peaks in intensity are retained while other peaks are deleted.

4A. For the remaining peaks of which the m/z values are below PrecursorIon, the intensity of each peak is amplified by five, while for peaks that are higher than PrecursorIon, the intensity is unchanged.

[B] Scoring of glycopeptide candidate against the processed spectrum:

1B. For an input *O*-glycopeptide composition, the *O*-glycan mass is added to the glycosylated residue.

2B. If ChargeState = 2, continue from 3B; if ChargeState \geq 3, continue from 4B.

3B. The monoisotopic m/z values of singly charged c-ions are calculated, starting from the c_1 ion, and the calculation is ended when the c-ion $>$ Spectrum[Maxima]. Count the total number, N , of the calculated singly charged c-ions only if:

$$\text{Spectrum[Minima]} \leq \text{singly charged c-ion} \leq \text{Spectrum[Maxima]}$$

One exception is that c-ions that are N-terminal to proline (P) residues are not counted towards the total number of c-ions.

The same calculation procedure is repeated for z-ions and count the total number of calculated singly-charged z-ions that conform to:

$$\text{Spectrum[Minima]} \leq \text{singly charged z-ion} \leq \text{Spectrum[Maxima]}$$

One exception is that z-ions of which the N-terminus are proline (P) residues are not counted towards the total number of z-ions.

Continue from 5B.

4B. The monoisotopic m/z values of singly charged c-ions are calculated, starting from the c_1 ion, and the calculation is ended when the c-ion $>$ Spectrum[Maxima]. Starting from the first singly charged c-ion that is above Spectrum[Maxima], calculate the corresponding doubly charged c^{2+} -ion, move to the next c^{2+} -ions along the sequence, and the calculation is ended when the doubly charged c^{2+} -ion $>$ Spectrum[Maxima]. Count the total number, N , of both the calculated singly charged c-ions and doubly charged c^{2+} -ions only if:

$$\text{Spectrum[Minima]} \leq \text{c-ion (c/c}^{2+}\text{-ion)} \leq \text{Spectrum[Maxima]}$$

One exception is that c-ions that are N-terminal to proline (P) residues are not counted towards the total number of c-ions.

The same calculation procedure is repeated for z-ions and count the total number of calculated singly-charged z-ions and doubly charged z^{2+} -ions that conform to:

$$\text{Spectrum[Minima]} \leq \text{z-ion (z/z}^{2+}\text{-ion)} \leq \text{Spectrum[Maxima]}$$

One exception is that z-ions of which the N-terminus are proline (P) residues are not counted towards the total number of z-ions.

Continue from 5B.

5B. The monoisotopic m/z values of singly charged y-ions are calculated, starting from y_1 ion, and the calculation is ended when the y-ion $>$ Spectrum[Maxima]. Count the total number of the calculated singly charged y-ions only if:

$$\text{Spectrum}[\text{Minima}] \leq \text{singly charged y-ion} \leq \text{Spectrum}[\text{Maxima}]$$

6B. For c-ion series, the calculated c-ions (including c^{2+} -ions in 4B if the step is taken) that conform to:

$$\text{Spectrum}[\text{Minima}] \leq \text{c-ion (c/c}^{2+}\text{-ion)} \leq \text{Spectrum}[\text{Maxima}]$$

are searched against the processed spectrum, and a match is counted only if:

$$(\text{Spectra}[m/z] - 1.2) \leq \text{Searched c-ion} \leq (\text{Spectra}[m/z] + 0.8)$$

The number of matched c-ions, n , is counted based on the number of matches. The intensity values of spectral peaks that are matched to c-ions are summed together.

The ion series score for c-ions is calculated based on the following formula:

$$\text{Score}(c - \text{ion}) = -10 \times \log \left[\sum_{k=n}^N \binom{N}{k} p^k (1-p)^{N-k} \right]$$

In the equation, N is the total number of theoretical c-ions searched, n is the number of matched c-ions, and p is the probability that a single ion is matched to the spectrum by chance.

Individual scores for z- and y-ion series are computed in the same way as c-ion series, except that for z-ion series, a match is counted only if:

$$(\text{Spectra}[m/z] - 1.7) \leq \text{Searched z-ion} \leq (\text{Spectra}[m/z] + 1.3)$$

In addition, the intensity values of peaks that are matched to z- and y-ions, respectively, are summed together.

7B. The weighting for each ion series (c-, z- and y-ions) is calculated by dividing the intensity of the matched peaks for that specific ion type by the total intensity of all matched peaks.

The total score of the input glycopeptide candidate is calculated by adding the products of the individual ion series' score and the corresponding weighting together, based on the following equation:

$$\text{Total Score} = \sum_{k=c, z, y} \left[\frac{\sum \text{Int. (k - ions)}}{\sum \text{Int. (all ions)}} \times \text{Score}(k - \text{ion}) \right]$$

8B. The value of the Total Score is reported for each input glycopeptide composition.

CHAPTER IV

Determination of the False Discovery Rate in N-Linked Glycopeptide Identifications by GlycoPep Evaluator

This work has been published by the journal Analytical Chemistry, with reprint permission from the journal.

Glycoproteins are biologically significant large molecules that participate in numerous cellular activities. In order to obtain site-specific protein glycosylation information, intact glycopeptides, with the glycan attached to the peptide sequence, are characterized by tandem mass spectrometry (MS/MS) methods such as collision-induced dissociation (CID) and electron transfer dissociation (ETD). While several emerging automated tools are developed, no consensus is present in the field about the best way to determine the reliability of the tools and/or provide the false discovery rate (FDR). A common approach to calculate FDRs for glycopeptide analysis, adopted from the target-decoy strategy in proteomics, employs a decoy database that is created based on the target protein sequence database. Nonetheless, this approach is not optimal in measuring the confidence of N-linked glycopeptide matches, because the glycopeptide data set is considerably smaller compared to that of peptides, and the requirement of a consensus sequence for N-glycosylation further limits the number of possible decoy glycopeptides tested in a database search. To address the need to accurately determine FDRs for automated glycopeptide assignments, we developed GlycoPep Evaluator (GPE), a tool that helps to measure FDRs in identifying glycopeptides without using a decoy database. GPE generates decoy glycopeptides *de novo* for every target glycopeptide, in a 1:20 target-to-decoy ratio. The decoys, along with target glycopeptides, are scored against the ETD data, from which FDRs can be calculated accurately based on the number of decoy matches and the ratio of the number of targets to decoys, for small

data sets. GPE is freely accessible for download and can work with any search engine that interprets N-glycopeptide ETD data.

4.1 Introduction

Glycosylation is commonly considered the most extensive post-translational modification on proteins, and it is estimated that 20%-50% of all proteins are glycoproteins.¹⁻² Glycosylation is known to impact protein folding and function;³⁻⁴ the interaction between proteins and glycans is a main route for cellular communications and signaling.⁵⁻⁷ In addition, changes in glycosylation pattern on certain proteins are closely related to the pathogenesis of diseases.⁸⁻⁹ Therefore, protein glycosylation analysis is a vital step towards understanding the role that carbohydrates play in various biological events.

One common method of characterizing the glycosylation on proteins is to digest the protein and to analyze the resulting glycopeptides. This strategy allows researchers to correlate the glycans to their attachment sites in the protein(s).¹⁰⁻¹² In glycopeptide analysis, the correct glycopeptide compositions usually cannot be determined by high resolution MS data alone, and MS/MS data are needed for confident glycopeptide assignments.¹³ In order to accelerate the analysis workflow for high-throughput glycopeptide identifications, an increasing number of bioinformatics tools are developed to analyze MS/MS data of glycopeptides.¹⁴⁻²⁰ Strum et al. presented a program called GlycoPeptide Finder that can interpret CID data of N- and O-glycopeptides generated from nonspecific proteolysis.²¹ A computational framework was developed to implement a software tool called GlycoFragwork, which is capable of scoring N-glycopeptide MS/MS data from multiple fragmentation modes.²² We recently introduced two web-based utilities, GlycoPep Grader²³ and GlycoPep Detector²⁴, to determine the most likely N-glycopeptide compositions by scoring the CID and ETD data against each of the possible

glycopeptide candidates. In all the applications described above, the glycopeptide analysis tool returns a best glycopeptide match for each MS/MS spectrum by selecting the candidate that receives the highest score under a certain scoring algorithm. Although these matches are very helpful in guiding the user, the top match is sometimes incorrect.

While automated analysis tools are helpful for glycopeptide analysis, users need to know the likelihood that the automated matches are correct. Therefore, it is important for any tool to provide users with a reliable false discovery rate (FDR), which is the measure of probability that a match is correct, based on the program's performance in analyzing the entire data set.²⁵⁻²⁸ The concept of calculating an FDR has been well established by the proteomics community, and to determine the FDR value in proteomics, a composite database is generated by combining the target protein sequence database and a decoy sequence database. The decoy database is nonsensical and created based on the target database such that they contain an equivalent number of peptide sequences, which is often accomplished by reversing the protein sequences in the target database.²⁶⁻³⁰ Subsequently, the MS/MS data are scored against the composite database, and the numbers of matches made against the target and decoy sequences are used to calculate FDR. Following the assumption that the distribution of incorrect matches to target sequences is the same as that of matches to decoy sequences, the number of false positive identifications, which directly translates to FDR, can be calculated by doubling the number of decoy matches. This target-decoy approach is simple and works well for peptide identifications based on large scale proteomics data.³¹⁻³³

Most of the currently available glycopeptide analysis tools do not have the capability to calculate FDRs for glycopeptide assignments, and for those that are enabled with this functionality, the target-decoy approach is adopted to estimate FDRs in glycopeptide identifications, where an equal amount of decoy glycopeptides are generated on the basis of the

target glycoprotein sequences to comprise the decoy database.²¹⁻²² However, in a glycoproteomics experiment, the number of CID or ETD spectra scored is considerably smaller than the number of spectra scored in a proteomics experiment. This is expected since glycoproteomics experiments are often conducted on a single protein, not thousands of proteins. Even when the entire proteome is evaluated for glycopeptides, the number of CID or ETD spectra that are verified to be from glycopeptides is generally much less than 1000. As a result, using the conventional approach for calculating FDRs, the distribution of decoy glycopeptide matches may not accurately reflect that of incorrect matches to target glycopeptides because the collected glycopeptide data set is not large enough.^{21, 34-35} Furthermore, for N-glycopeptides, a consensus sequence of N-X-S/T (X can be any amino acid except proline) must be present, which further limits the number of possible decoy glycopeptides being tested. All these factors lead to inaccurate FDRs when the target-decoy approach is applied to small to moderate size glycoproteomics data sets.

In this work, we present a new method to determine FDRs with high accuracy for N-linked glycopeptide identifications based on ETD data. Instead of creating a decoy database of the same size as the target database, we developed a tool called GlycoPep Evaluator (GPE) to generate decoy glycopeptides *de novo* for every target glycopeptide, in a 1:20 target-to-decoy ratio. The decoys are made under specific rules so that they contain the consensus sequence for N-glycosylation, while they have distinct glycopeptide sequences and glycosylation sites. To determine the FDR, all the generated decoys are scored against the ETD data along with target glycopeptides, and the FDR is calculated accurately based on the number of decoy glycopeptide matches and the relative amount of targets to decoys. GPE is freely available for download and can be used in conjunction with any scoring schemes for assessing ETD data of glycopeptides.

4.2 Experimental Procedures

4.2.1 Samples and Reagents

Bovine fetuin, RNase B, and human serum proteins (IgG, AGP, transferrin) were obtained from Sigma Aldrich (St. Louis, MO). The HIV envelope protein, C.97ZA012 gp140, was provided by the Duke Human Vaccine Research Institute (Durham, NC).³⁶ Sequencing grade trypsin was purchased from Promega (Madison, WI). All chemical reagents used were either of analytical grade or better.

4.2.2 Protease Digestion

Glycoproteins of 72-100 µg were dissolved in 100 mM Tris buffer at pH 8 with a concentration of 2.4-3.3 µg/µL. Samples were denatured by addition of solid urea so that the urea concentration was 6 M, followed by addition of 5 mM tris(2-carboxyethyl)-phosphine (TCEP) solution to reduce the disulfide bonds (the molar ratio of TCEP to disulfide bond was kept at 6:1), and 10 mM iodoacetamide (IAM) was subsequently added to alkylate the free thiol groups using a molar ratio of 8:1. The reaction was left to proceed for 1 h at room temperature in the dark. Dithiothreitol (DTT) solution was then added to a final concentration of 10 mM to quench the alkylation reaction. Prior to enzymatic digestion, the urea concentration was decreased to 1 M by diluting the samples with Tris buffer. Subsequently, trypsin was added at a 1:30 enzyme-to-protein ratio, followed by 18 h incubation of the samples at 37 °C. Finally, trypsin digestion was stopped by adding 1 µL acetic acid for every 100 µL of glycoprotein solution. The prepared samples were stored at -20 °C before subjected to LC/MS analysis.

4.2.3 LC/MS Analysis

Digested glycoprotein samples were analyzed using a Waters Acquity Ultra Performance Liquid Chromatography system (Milford, MA) coupled to a LTQ Velos linear ion trap mass spectrometer (Thermo Scientific, San Jose, CA). For each run, 5 μL of a sample was injected onto a capillary C_{18} column (300 μm i.d. \times 5 cm, 100 \AA , Micro-Tech Scientific, Vista, CA). Two mobile phases were employed for separation: solvent A consists of 99.9% H_2O plus 0.1% formic acid, and solvent B consists of 99.9% acetonitrile with 0.1% formic acid. The LC separation gradient was as follows: 2% solvent B for 5 min, followed by a linear increase to 40% B in 50 min, and a ramp to 90% B in 10 min.³⁷⁻³⁸ The column was kept at 90% solvent B for an additional 10 min and then re-equilibrated at 2% B for 10 min. The mass spectrometer was operated in the positive ion mode, with the ESI source voltage at 3 kV and the capillary temperature set at 200 $^{\circ}\text{C}$. For the data-dependent acquisition, CID and ETD spectra were collected by selecting the five most intense peaks in the full scan MS (m/z 500-2000) and the precursor ions were fragmented in either CID or ETD mode. The normalized collision energy was at 30% with an activation time of 10 ms for CID, and the ion-ion reaction time was set at 90 ms for ETD with supplemental activation turned on. The automatic gain control (AGC) target value was set at 2×10^4 for MS/MS scans in the ion trap, and the AGC target value of the fluoranthene reagent anions was at 2×10^5 . The reaction time between anions and cations in ETD was set at 90 ms, and the supplemental activation was turned on for ETD so that precursor ions and charge-reduced species could undergo further dissociation. For CID, the normalized collision energy was set at 30%, with activation time of 10 ms.

4.2.4 Glycopeptide MS/MS Data Set

In this study, MS/MS data were collected on glycoproteins that have been previously characterized in the literature.^{36, 40-42} *In silico* trypsin digestion was performed on the glycoprotein sequences with up to 2 missed cleavages allowed, and carbamidomethylation was set as a fixed

modification on cysteine residues. Theoretical monoisotopic masses of potential N-glycopeptides were calculated by adding the site-specific N-glycan masses to the masses of the corresponding peptides that contain the glycosylation sites. The theoretical m/z values of these glycopeptides were then computed and searched against the ETD data to see whether precursor ions of these m/z values were selected for ETD. Manual analysis was then performed on every identified ETD spectrum that may come from potential glycopeptides. If a match was found, CID data were employed to further confirm the glycopeptide assignment. In this way, a glycopeptide ETD data set with known glycopeptide compositions was built that includes glycopeptides of diverse peptide sequences and varying glycan types.

4.2.5 Decoy and Target Candidates Generation

For this study, all of the glycopeptide assignments were known. However, to demonstrate our approach, we simulated a case where the identity of the glycopeptide was not known, and the user had to choose between multiple feasible candidates. Therefore, we needed mock candidates and decoys to score against each spectrum. GlycoPep Evaluator (GPE) was used to generate 20 decoys per candidate. The correct “candidate” for each spectrum is known, and the additional mock candidates were generated using GlycoMod.⁴² To generate the mock candidates, sequences of the studied glycoproteins were entered into GlycoMod, along with a polypeptide sequence, Titin, which contains 50,000 amino acid residues. As a result, multiple glycopeptide compositions were produced by GlycoMod for every glycopeptide peak that was subjected to ETD (with a mass tolerance of 200 ppm), and a selection of the glycopeptides were entered into GPE as (mock) target glycopeptide candidates. Typically, five candidate glycopeptides were entered, where one of the candidates was the true glycopeptide. For each target glycopeptide, GPE is used to generate 20 decoy glycopeptides of isobaric masses, and these decoys can be used for evaluating the false

discovery rate (FDR) in automated assignment of glycopeptides by a search engine. GPE includes functionality to generate any number of decoys, but 20 were used herein.

4.2.6 Scoring of Decoy and Target Candidates

GPE is a freely available software tool that we developed to determine FDRs in glycopeptide analysis. Apart from the function to generate decoy glycopeptides, GPE also incorporates an ETD algorithm that we described previously,²⁴ and it can score each target and decoy candidate against the ETD spectrum in an automated manner. In order to use the scoring functionality of GPE, the user needs to upload the raw ETD data, specify the MS/MS scan range and the ion types being scored, and submit the target and decoy candidates for scoring. GPE will then generate the result page where the candidates are ranked by the scores that they are assigned.

4.2.7 False Discovery Rate Study Using GPE

To demonstrate how to obtain an accurate FDR using GlycoPep Evaluator, GPE was enabled to score glycopeptide ETD data by itself using the algorithm that we developed before.²⁴ To use the scoring function of GPE, a raw ETD spectrum was first converted to a .txt file that contains the m/z values and peak intensities, and the file was then uploaded into GPE directly. GPE scores the ETD data against all the generated decoy glycopeptides as well as the target glycopeptides, and a target or a decoy match is made depending on whether a target or a decoy candidate receives the highest score. Using the number of decoy matches made by GPE in assessing the glycopeptide data set and the target-to-decoy ratio (1:20 in our study), the FDRs in glycopeptide analysis could be calculated.

4.3 Results and Discussion

4.3.1 Overview of GlycoPep Evaluator

GlycoPep Evaluator (GPE) is a freely downloadable software tool that can be used to generate decoy glycopeptides for false discovery rate analysis. It also has incorporated functionality to score all the targets and decoys against imported spectra using a previously published scoring algorithm.²⁴ GPE was written in Java and developed with Java Development Kit 7 (JDK 7). The program has been tested to perform successfully under Windows and Linux systems, and Java Runtime Environment 7 (JRE 7) is recommended to be installed prior to running GPE.

The graphical user interface (GUI) of GPE is shown in Figure 1A. To generate decoy glycopeptides, the user needs to enter the target glycopeptide sequence and to specify the N-glycosylation site location by entering the Glycosylated Asn Index (if a default value of 0 is input, the software will automatically locate the first Asn that meets the N-X-S/T sequon). Cysteine modifications can be selected by the user as indicated in the GUI; if there is an additional modification on any amino acid residue, the user can specify the location and the mass of the modification as needed. For the glycan portion, the user can either type in the number of each monosaccharide unit (Hex, HexNAc, Neu5Ac, etc.) that constitutes the glycan or input the glycan mass, as shown in Figure 1A. Other parameters that are necessary to generate decoys include the precursor ion's m/z and charge state, mass tolerance (in ppm), number of maximum missed cleavages, peptide variation (in Da, see discussion below) and the number of decoys per target. The mass tolerance is the mass range that the monoisotopic mass of a decoy glycopeptide, as generated by GPE, is allowed to deviate from the precursor ion's mass (as calculated by the precursor ion's m/z and charge state). The peptide variation, on the other hand, is the mass range that the peptide portion of the decoy (calculated by subtracting the glycan mass from the monoisotopic mass) is allowed to differ from that of the peptide in the target glycopeptide. In our

experiments, the mass tolerance for decoys was set at 20 ppm, maximum missed cleavage number was set to 2, peptide variation was set at 200 Da and the number of decoys per target was set to 20. Currently, the tool specifically generates tryptic peptides. If sufficient interest warrants future development, other options for peptide generation could be included.

Once the required parameters are submitted to generate decoy glycopeptides, GPE will present the result page where 20 output decoys are listed, as exemplified in Figure 1B. Several requirements are met by GPE in producing the decoy glycopeptide candidates: First, the decoy ends with either Arg or Lys on its C-terminus; second, the missed cleavages on the decoy sequence must not exceed the number of maximum missed cleavages specified by the user; third, the decoy contains a consensus sequence, Asn-X-Ser/Thr (X is any random amino acid, excluding proline), with the Asn being the glycosylation site; fourth, the peptide portion of the decoy has a mass that is within a user-specified range (termed “peptide variation”) from the peptide mass of the target glycopeptide; finally, the glycan portion of the decoy is assigned a mass that makes the m/z of the entire decoy within the user-specified mass tolerance of the precursor ion’s m/z , and the glycan mass value is appended to the glycosylated Asn as a modification of mass in the output of the decoy glycopeptide (Figure 1B).

GlycoPep Evaluator

Gen Decoys **Result**

A

Glycopeptide

A.A.sequence: LTPLCVTLHCTNATFK

Cys Modification: ☐ None ☒ Iodoacetamide ☐ Iodoacetic acid ☐ 4-vinyl pyridine

Glycosylated Asn Index: Mod. Location: Mod. Mass:

☒ Glycan composition ☐ Glycan Mass

[Hex]: [HexNAc]: [Neu5Ac]:
[Fuc]: [NeuGc]: [SO3]: Glycan Mass:

Parameters for decoys and scoring

Precursor m/z: Charge state:
Mass tolerance: Missed cleavages:
Peptide variation: Number of decoys:
Scan lower bound: Scan upper bound:

☐ c-ion ☐ z-ion ☐ y-ion ☐ c2+-ion ☐ z2+-ion

GlycoPep Evaluator

Gen Decoys **Result**

B

info	Glycopeptide composition	Monoisotopic m/z
target	LTPLCVTLHCTN(1647.6133)ATFK	1175.1912
decoy	N(1670.5888)LTWVNKTQAHKQER	1175.1937
decoy	WPNCKFRTAGN(1768.6493)FTR	1175.1728
decoy	WQQTEKEGSRNQN(1462.5797)ATAR	1175.1933
decoy	RSQTGNFLSN(1758.7736)ESCHK	1175.1988
decoy	EFTMN(1542.6843)HSVKHPQPWDK	1175.2125
decoy	WLN(1835.7743)MSWPGSPQDEK	1175.1929
decoy	YRGEHKMPWPDQN(1490.6055)CSK	1175.1773
decoy	SFYTHAQCN(1614.6694)GTHQTTR	1175.1749
decoy	RQFN(1453.5795)DSWMPAGMSRCVK	1175.1803
decoy	SAYWN(1576.7873)VTQRCWCCR	1175.2076
decoy	TVSN(1493.4781)QTKFNRFTCWLK	1175.1734
decoy	VLN(1718.6526)WTGVRCTAPQFR	1175.1983
decoy	AANFHCWMYN(1526.6530)LTKANR	1175.1982
decoy	YLDNEN(1575.7259)QTHHGDYWR	1175.1925
decoy	HAYMMSYVYTHN(1593.7334)RTR	1175.2109
decoy	QMN(1632.7438)DTS DGWYGHNPLR	1175.1924
decoy	VDGQN(1575.6786)TTHGPYTDVQCR	1175.1849
decoy	QSQVKGVALN(1650.5944)TSPDTLSK	1175.2039
decoy	TERNRGN(1705.6406)CSPVLMQR	1175.1803
decoy	QWHKVPWN(1546.6028)ESLYHPR	1175.2017

Figure 1. (A) The graphical user interface (GUI) of the GlycoPep Evaluator (GPE) program. (B) The result of decoy generation completed by GPE that contains the input target glycopeptide as well as 20 decoy glycopeptide sequences generated by the program.

Following these rules, the generated decoy glycopeptide can closely mimic the target glycopeptide in terms of the glycosylation site, protease specificity and the approximate peptide length. On the other hand, 20 decoy glycopeptides of distinct sequences and varying glycan locations are produced for every single target glycopeptide, as demonstrated in Figure 1B, thus providing a sufficient number of decoy candidates that can compete with the target glycopeptides in the scoring by a software tool.

4.3.2 False Discovery Rate Analysis

The false discovery rate (FDR) is, by definition, the percentage of accepted peptide-spectral matches that are incorrect.²⁸ When decoys are included in database searching, the incorrect matches are comprised of a proportion of the target matches as well as all the decoy matches. The latter are used to estimate the number of target matches that are incorrect. As such, FDR is calculated by the following equation:

$$\text{FDR} = \frac{N_{\text{ic}} + N_{\text{d}}}{\text{Total Assignments}} \quad (1)$$

In the equation, N_{ic} is the number of incorrect assignments made to target candidates and N_{d} is the number of decoy assignments.

Because both the incorrect target matches and the decoy matches are made at random, the number of hits for incorrect target assignments or decoy assignments is proportional to the number of the corresponding target or decoy candidates scored by a program. Consequently, the ratio of the number of incorrect target assignments to decoy assignments is equal to the ratio of target candidates to decoy candidates in quantity:

$$\frac{N_{\text{ic}}}{N_{\text{d}}} = \frac{\text{Number of Targets}}{\text{Number of Decoys}} \quad (2)$$

When equations 1 and 2 are combined, the FDR is determined by equation 3:

$$\text{FDR} = \frac{N_d}{\text{Total Assignments}} \times \left(1 + \frac{\text{Number of Targets}}{\text{Number of Decoys}}\right) \quad (3)$$

In a conventional workflow, since an equal number of decoy sequences are scored along with target sequences, N_{ic}/N_d is 1. Therefore, according to equation 3, the FDR is calculated by doubling the number of decoy matches divided by the number of total assignments. In our method, however, the target-to-decoy ratio is 1:20 rather than 1:1 because 20 decoy candidates are generated and scored for each target, thus N_{ic}/N_d is 0.05. Accordingly, FDR is determined by equation 4:

$$\text{FDR} = \frac{N_d}{\text{Total Assignments}} \times 1.05 \quad (4)$$

Consequently, using our method in which 20 decoy glycopeptides are created and tested for every target glycopeptide composition, the FDR can be measured based on the number of decoy matches and the number of total accepted assignments, as formulated in equation 4.

4.3.3 Target and Decoy Glycopeptides Analysis

Apart from generating decoy glycopeptide candidates *de novo*, GPE was also implemented with an algorithm that we developed to process and score ETD data of N-glycopeptides.²⁴ After a list of decoy candidates are generated by GPE, the user can load raw ETD data to the program and specify the MS/MS scan range; GPE can score all the decoy candidates as well as the target glycopeptide compositions against the input MS/MS data. For every glycopeptide composition, GPE evaluates the match of different ion series (c, z and y-ions) to the processed ETD data and assigns a final score to each candidate, as described in the algorithm published with Reference 24.

The decoy glycopeptides can then be sorted from high score to low and be compared with the scores of target glycopeptides.

To demonstrate the functionality of GPE, we present, below, a CID and ETD spectrum of a known glycopeptide and show how the GPE would process the ETD data, score the spectrum, and then additionally calculate scores for decoy assignments. Figure 2A is the ETD data of a glycopeptide from HIV gp140 that has a composition of DGGEDNKTEEIFRPGGGNMK + [Hex]3[HexNAc]4[Fuc]1 (where *N* is the glycosylation site). In the ETD spectrum, c-ions (c_4 - c_5) and z-ions (z_8 - z_{13}) are observed that can be used to determine the glycopeptide's sequence, as shown in the figure. Additionally, the CID data in Figure 2B further confirms that the precursor ion is a glycopeptide peak, because glycan oxonium ions are present at m/z 366 and 528. Moreover, monosaccharide losses, including losses of Hex, HexNAc and glycan dissociation patterns in CID, the glycan portion of the glycopeptide can be deduced to be [Hex]3[HexNAc]4[Fuc]1. It is noteworthy that, although CID data are utilized to verify glycopeptide assignments, in our method, we did not implement CID fragmentation rules in the scoring function, and only ETD data should be submitted to GPE for appropriate FDR analysis.

To demonstrate that the glycopeptide composition described above can be correctly assigned by GPE, the true glycopeptide composition, along with four isobaric glycopeptide “mock” candidates, were entered into GPE as potential target glycopeptide candidates. GPE then generated 20 decoy glycopeptides per target. The ETD data were subsequently submitted to the software, and all the candidates (including decoys) were scored by GPE. A total of 100 decoy glycopeptide compositions were created by GPE for the 5 target glycopeptides, and each decoy has its distinct sequence and N-glycosylation site. The true glycopeptide composition, its 20 decoys, and the associated scores are shown in Figure 3; the remaining 4 targets, their 80 decoys, and their scores are shown in Table S1. The correct glycopeptide composition, labeled as target in Figure 3,

receives the highest score of 61.7, which is significantly higher than the score of any other candidate, including the other 4 target and 100 decoy glycopeptides. By contrast, none of the other 4 input target glycopeptides (which are incorrect candidates but still considered “targets”, for the purposes of this demonstration), outscore the best decoy glycopeptide sequences generated by GPE. While at least one of the 20 decoys in each of these sets outscore the falsely generated “target” candidates, the overall highest scoring decoy, with a score of 17.8, does not outscore the true assignment. (Additional data were shown in Figure S1.) Therefore, the first glycopeptide candidate, which is also the manually verified correct assignment, is assigned to the ETD data by GPE, even when four other incorrect candidates and 100 decoys are scored in parallel. This example shows how to use GPE: correct and incorrect target glycopeptides can be readily differentiated by including a sufficient number of decoy glycopeptides in the scoring process, which are generated by GPE in an automated fashion.

Name	Composition	Mono <i>m/z</i>	Score
target	DGGEDN(1444.5339)KTEEIFRPGGGNMK	1199.1752	61.6754
decoy 1	N(1588.5408)VSCKKPVWCMPSGGGSR	1199.1637	8.8906
decoy 2	YDN(1588.6667)RSMRGMFGFTHCK	1199.1865	8.7011
decoy 3	AKYPTGN(1380.4808)DTFENESDVKGSR	1199.1739	6.2219
decoy 4	YFWCPKRWQN(1667.5246)LSSR	1199.1592	5.9449
decoy 5	LN(1437.3562)LTWSKHTHNRLPTDK	1199.1741	5.4633
decoy 6	VFVLVGNATN(1413.4413)MSESWCDPR	1199.1538	4.5258
decoy 7	N(1158.3110)MTKWPNEHWRNGMLHMVR	1199.1614	4.3031
decoy 8	PEADMQGSREPN(1461.4535)ATFWAVK	1199.1568	4.1652
decoy 9	YYYMGNLVMNN(1127.3184)YTPVQWRR	1199.1656	4.0256
decoy 10	GYEWN(1592.6778)KSGMGEMHWYK	1199.1833	3.8474
decoy 11	GFVMHSFTNSPCN(1620.7062)DSFK	1199.1894	3.8474
decoy 12	FNDWALMN(1287.4151)QSVNMVWPTVR	1199.1766	3.5206
decoy 13	FQFFN(1316.3959)DSRKNTHQFPHTK	1199.1737	3.3116
decoy 14	LNASPQKN(1537.3879)HTYWVSTRR	1199.1551	3.2378
decoy 15	YVDRTSYDYSRSHDN(1288.5189)WSR	1199.1836	3.0834
decoy 16	PGRTCETDGHVTPLAN(1485.4418)QTR	1199.1553	3.0834
decoy 17	N(1412.5759)YTAFESCMTNMLWKMR	1199.1816	2.8765
decoy 18	DCYRYCPYKN(1562.6731)SSEYK	1199.1802	0
decoy 19	VQDWWTN(1327.4822)ATMNVALNYWR	1199.1874	0
decoy 20	DCWEYN(1395.4918)ATLKNNDKDWK	1199.1609	0

Figure 3. For the input glycopeptide composition (labeled as target) DGGEDN/KTEEIFRPGGGNMK + [Hex]3[HexNAc]4[Fuc]1, 20 decoy glycopeptide compositions were generated by GPE. Subsequently, GPE scored both the target and decoy glycopeptides against the ETD data, and they were ranked from high to low score as shown in this figure. The target glycopeptide, which is also the correct assignment, received the highest score of 61.7, outscoring all the other candidates. The scoring results of the other four incorrect glycopeptide candidates are summarized in Figure S1.

4.3.4 Is GPE Consistently Able To Identify The Correct Candidate, When It Is Present?

The above example illustrates that GPE can be used to effectively identify a correct target candidate among a large list of incorrect glycopeptides. To determine how consistently GPE could generate these kinds of successful results, we tested a larger data set. We employed GPE in analyzing a glycopeptide data set that contains ETD data of 77 distinct glycopeptides generated

from multiple proteins (fetuin, IgG, HIV gp140, etc.). In these cases, all 77 spectra were manually assigned using the same procedure described above. After determining the correct assignment for each spectrum, 4 other (incorrect) “target” assignments were also generated. The software assigned 76 of the 77 MS/MS spectra to the correct glycopeptide compositions, demonstrating that the approach can consistently return the correct result, even when 20 decoys per candidate are scored. These results are expected when a high-quality algorithm is used for scoring glycopeptides, such as the one used in GPE, and the spectra are of high enough quality such that manual assignment is possible.

4.3.5 Is GPE Effective At Identifying Misassigned Spectra?

We next tested whether GPE is capable of indicating that the incorrect target glycopeptides are incorrect when the true candidates are not present in the target list. The correct glycoprotein sequences that generated the ETD data were excluded from the search of target glycopeptide compositions, so that all the target glycopeptides were incorrect glycopeptide candidates from Titin. After the incorrect targets were input into GPE, they were scored along with 20 decoys per target. Only four out of the 77 ETD spectra were matched to the target glycopeptides that are incorrect, whereas 73 spectra were assigned to decoy glycopeptides. Therefore, the ratio of incorrect target matches to decoy matches, N_{ic}/N_d , is 0.055 (4/73) in this case. This value is very close to the target-to-decoy ratio of 0.05 (1/20).

4.3.6 Comparison of the predicted FDR to the true FDR

Using the data above, we evaluated the true FDR for our data set of 77 spectra compared to the FDR that would be predicted by equation 4. When the correct glycopeptide compositions are included in the test, as mentioned above, 1 out of 77 assignments is a decoy match, and the FDR,

according to equation 4, is predicted to be 1.36%. The actual FDR that is observed, on the other hand, is the number of incorrect assignments divided by the total assignments. In this case, only the decoy assignment is incorrect and the other 76 assignments are correct, so the observed FDR is 1.30% (1/77), which is closely approximated by the predicted FDR value. On the other hand, when the correct glycopeptide sequences are excluded from the target list, 73 of 77 assignments are decoy matches, which leads to a calculated FDR of 99.55%. (This calculation is done using Equation 4: $(73/77) \times 1.05 = 0.995$) The actual FDR is 100% since all the assignments are incorrect. In both circumstances, the predicted FDRs are very close to the observed FDRs.

To further test if FDR values for small data sets can be accurately determined by our method, a proportion of the 77 ETD spectra were randomly selected, and for those spectra, the correct glycoprotein sequences were excluded for generating target glycopeptide candidates. For the remaining spectra, the correct glycoproteins were included in the generation of target compositions. Subsequently, GPE was employed to score each ETD spectrum against the corresponding target glycopeptides, and the number of decoy assignments was used to calculate FDR based on equation 4. The experiment was conducted at 12 different cases such that 0, 3, 5, 10, 20, 30, 40, 50, 60, 70, 73, 77 out of the 77 correct glycopeptide sequences were randomly excluded when their respective spectra were being scored. In this way, different numbers of incorrect assignments for the ETD data set were generated, and the predicted FDR using our method can be compared to the observed FDR at different levels. The comparison of the calculated versus observed FDRs for the 77 tested ETD spectra is illustrated in Figure 4A, where a correlation curve is made based on the blue data points. The least-squares fitting line has a slope that only deviates slightly from unity, and the curve has good linearity (R^2 above 0.99). These data demonstrate that for glycopeptide data set with a wide range of FDRs (ranging from 1.3%-100%), the FDR values can be determined accurately using GPE and the method that we developed.

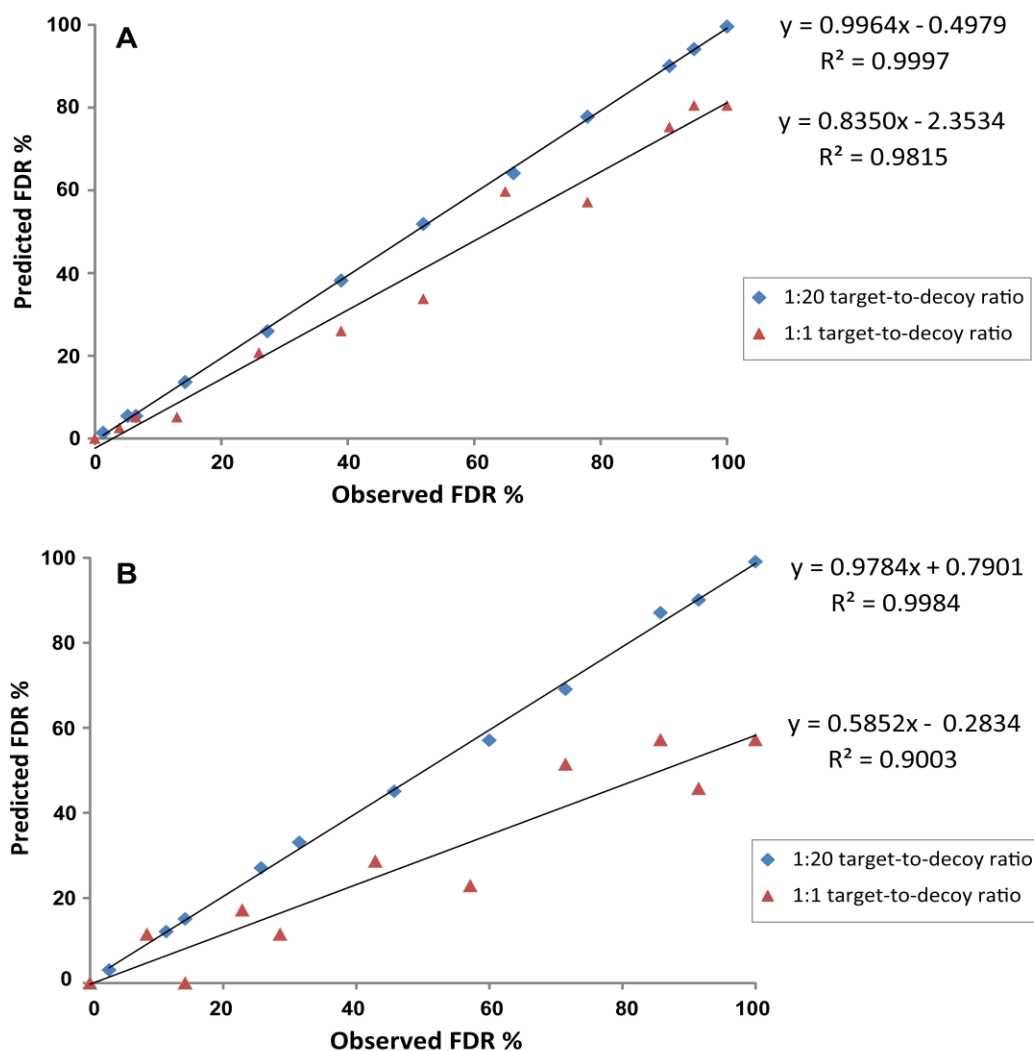


Figure 4. Lines that are fitted based on the blue data points: correlation curves between the predicted FDR values calculated using our method, and the observed FDR values that are manually verified. Lines that are fitted based on the red data points: correlation curves between the FDRs calculated by the common approach where an equal number of decoys are tested with the targets, and the observed FDRs that are verified manually. The FDRs are based on the analysis of ETD data sets of (A) 77 distinct glycopeptides and (B) 35 distinct glycopeptides using GPE program.

In glycopeptide-based identifications, the MS/MS data set is frequently of a small size, and a robust method needs to be able to determine the FDRs for these types of data. To build a smaller glycopeptide data set, we randomly selected 35 ETD spectra from the entire data set, and performed the same experiment as described above, to test whether using our method, the FDRs at

different levels can be measured with high accuracy for this data set that has a limited size. The result is shown in Figure 4B where the correlation curve is fitted based on the blue data spots; the best-fitting line between the predicted and observed FDRs has a slope that is close to 1 with R^2 still above 0.99. Therefore, these experiments prove that the developed method is accurate in measuring the FDRs in glycopeptide identifications, even for glycopeptide data sets of small sizes.

Finally, the accuracy of our method in predicting the FDRs was compared to that of the common approach where an equal number of decoy glycopeptides were tested with the target glycopeptides. For the same two data sets described above, the correlation curves between the FDRs predicted by the latter approach with 1:1 target-to-decoy ratio and the observed FDRs, are also shown in Figure 4. In this experiment, an equal number of decoy glycopeptides were generated by GPE based on the target candidates, and both the decoy and target glycopeptides are analyzed in the same way as described previously. These data sets are present in red. For the 77 tested ETD spectra, the R^2 of the curve is below 0.99, and the slope of the curve (0.83) deviates from 1 (Figure 4A). Furthermore, using the conventional approach, the correlation between the predicted and observed FDRs becomes much worse when the size of the data set decreases, as evidenced by the correlation curve in Figure 4B that has a R^2 of only 0.90 and a flat slope of 0.58. The slope of the curves reflect the ratio of predicted FDRs to true FDRs, and a significantly less value than 1 indicate that true false positives would be considerably underestimated using the conventional approach. By contrast, the FDRs are predicted accurately using our method, especially under circumstances where only a small glycopeptide data set is available.

4.4 Conclusions

False discovery rate (FDR) is an important measurement of the confidence of glycopeptide assignments when MS/MS data of glycopeptides are analyzed. In order to accurately determine the

FDR of glycopeptide identifications, we developed a software program, GlycoPep Evaluator, to generate abundant decoy glycopeptide compositions and to score the target and decoy glycopeptide candidates in measuring the FDR. The target-to-decoy ratio is 1:20 so that even for a small number of target glycopeptide sequences, sufficient decoy glycopeptides are available for scoring, hence false-positive identifications can be better contained. Moreover, FDRs can be measured with high accuracy using GPE for small data sets, which are commonly seen in glycoproteomics where tens to hundreds of spectra are scored, as opposed to thousands of spectra scored in a proteomics experiment. The capability of GPE in generation of decoy glycopeptide candidates can be combined with any other data analysis tools that score ETD data of glycopeptides, so that FDRs can be accurately determined.

References

- (1) Apweiler, R.; Hermjakob, H.; Sharon, N. *Biochim. Biophys. Acta-Gen. Subj.* **1999**, *1473*, 4-8.
- (2) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. *Scientific Reports.* **2011**, *1*.
- (3) Banerjee, S.; Vishwanath, P.; Cui, J.; Kelleher, D. J.; Gilmore, R.; Robbins, P. W.; Samuelson, J. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 11676-11681.
- (4) Banks, D. D. *J. Mol. Biol.* **2011**, *412*, 536-550.
- (5) Furukawa, K.; Ohkawa, Y.; Yamauchi, Y.; Hamamura, K.; Ohmi, Y. *J. Biochem.* **2012**, *151*, 573-578.
- (6) Van den Steen, P.; Rudd, P. M.; Dwek, R. A.; Opdenakker, G. *Crit. Rev. Biochem. Mol. Biol.* **1998**, *33*, 151-208.
- (7) Coelho, V.; Krysov, S.; Ghaemmighami, A. M.; Emara, M.; Potter, K. N.; Johnson, P.; Packham, G.; Martinez-Pomares, L.; Stevenson, F. K. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18587-18592.
- (8) Wang, Y.; Tan, J.; Sutton-Smith, M.; Ditto, D.; Panico, M.; Campbell, R. M.; Varki, N. M.; Long, J. M.; Jaeken, J.; Levinson, S. R.; Wynshaw-Boris, A.; Morris, H. R.; Le, D.; Dell, A.; Schachter, H.; Marth, J. D. *Glycobiology.* **2001**, *11*, 1051-1070.
- (9) Li, Y.; Tian, Y. A.; Rezai, T.; Prakash, A.; Lopez, M. F.; Chan, D. W.; Zhang, H. *Anal. Chem.* **2011**, *83*, 240-245.
- (10) Leymarie, N.; Zaia, J. *Anal. Chem.* **2012**, *84*, 3040-3048.
- (11) Desaire, H. *Mol. Cell. Proteomics.* **2013**, *12*, 893-901.
- (12) Hong, Q.; Lebrilla, C. B.; Miyamoto, S.; Ruhaak, L. R. *Anal. Chem.* **2013**, *85*, 8585-93.
- (13) Desaire, H.; Hua, D. *Int. J. Mass Spectrom.* **2009**, *287*, 21-26.
- (14) Woodin, C. L.; Maxon, M.; Desaire, H. *Analyst.* **2013**, *138*, 2793-2803.

- (15) Ceroni, A.; Maass, K.; Geyer, H.; Geyer, R.; Dell, A.; Haslam, S. M. *J. Proteome Res.* **2008**, *7*, 1650-1659.
- (16) Premier Biosoft. <http://www.premierbiosoft.com/glycan/index.html> (accessed April 30, 2014).
- (17) Goldberg, D.; Bern, M.; Parry, S.; Sutton-Smith, M.; Panico, M.; Morris, H. R.; Dell, A. *J. Proteome Res.* **2007**, *6*, 3995-4005.
- (18) Apte, A.; Meitei, N. S. *Methods in molecular biology (Clifton, N.J.)*. **2010**, *600*, 269-81.
- (19) Chandler, K. B.; Pompach, P.; Goldman, R.; Edwards, N. *J. Proteome Res.* **2013**, *12*, 3652-3666.
- (20) Bruker Daltonics. <http://glycomics.ccruc.uga.edu/GlycomicsPortal/showEntry.action?id=116> (accessed April 30, 2014).
- (21) Strum, J. S.; Nwosu, C. C.; Hua, S.; Kronewitter, S. R.; Seipert, R. R.; Bachelor, R. J.; An, H. J.; Lebrilla, C. B. *Anal. Chem.* **2013**, *85*, 5666-5675.
- (22) Mayampurath, A.; Yu, C. Y.; Song, E. W.; Balan, J.; Mechref, Y.; Tang, H. X. *Anal. Chem.* **2014**, *86*, 453-463.
- (23) Woodin, C. L.; Hua, D.; Maxon, M.; Rebecchi, K. R.; Go, E. P.; Desaire, H. *Anal. Chem.* **2012**, *84*, 4821-4829.
- (24) Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 5023-5032.
- (25) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. *Nat. Methods.* **2007**, *4*, 787-797.
- (26) Elias, J. E.; Gygi, S. P. *Nat. Methods.* **2007**, *4*, 207-214.
- (27) Choi, H.; Nesvizhskii, A. I. *J. Proteome Res.* **2008**, *7*, 47-50.
- (28) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *J. Proteome Res.* **2008**, *7*, 29-34.
- (29) Nesvizhskii, A. I. *Journal of Proteomics.* **2010**, *73*, 2092-2123.
- (30) Huttlin, E. L.; Hegeman, A. D.; Harms, A. C.; Sussman, M. R. *J. Proteome Res.* **2006**, *6*, 392-398.

- (31) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. *Mol. Cell. Proteomics*. **2009**, *8*, 2405-2417.
- (32) Reidegeld, K. A.; Eisenacher, M.; Kohl, M.; Chamrad, D.; Korting, G.; Blueggel, M.; Meyer, H. E.; Stephan, C. *PROTEOMICS*. **2008**, *8*, 1129-1137.
- (33) Wang, G.; Wu, W. W.; Zhang, Z.; Masilamani, S.; Shen, R. F. *Anal. Chem.* **2009**, *81*, 146-159.
- (34) Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P. A. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111-1120.
- (35) Chalkley, R. J. *J. Proteome Res.* **2013**, *12*, 1062-1064.
- (36) Go, E. P.; Chang, Q.; Liao, H. X.; Sutherland, L. L.; Alam, S. M.; Haynes, B. F.; Desaire, H. *J. Proteome Res.* **2009**, *8*, 4231-4242.
- (37) Zhu, Z.; Su, X.; Clark, D.F.; Go, E.P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 8403-8411.
- (38) Zhu, Z.; Go, E.P.; Desaire, H. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 1012-1017.
- (39) Alley, W. R.; Mechref, Y.; Novotny, M. V. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 161-170.
- (40) Wada, Y.; Azadi, P.; Costello, C. E.; Dell, A.; Dwek, R. A.; Geyer, H.; Geyer, R.; Kakehi, K.; Karlsson, N. G.; Kato, K.; Kawasaki, N.; Khoo, K. H.; Kim, S.; Kondo, A.; Lattova, E.; Mechref, Y.; Miyoshi, E.; Nakamura, K.; Narimatsu, H.; Novotny, M. V.; Packer, N. H.; Perreault, H.; Peter-Katalinic, J.; Pohlentz, G.; Reinhold, V. N.; Rudd, P. M.; Suzuki, A.; Taniguchi, N. *Glycobiology*. **2007**, *17*, 411-422.
- (41) Zhang, Y.; Go, E. P.; Desaire, H. *Anal. Chem.* **2008**, *80*, 3144-3158.
- (42) Cooper, C. A.; Gasteiger, E.; Packer, N. H. *Proteomics*. **2001**, *1*, 340-349.

Name	Composition	Mono <i>m/z</i>	Score	Name	Composition	Mono <i>m/z</i>	Score
target	NAAGN(1743.6078)FSEPSDSSGAITAR	1199.1564	7.1353	target	WEPPLDDGGSEIIN(1419.5022)YTLEK	1199.1871	6.4289
decoy	ARMEARFN(1691.6261)FTEFGAEK	1199.1842	13.9909	decoy	SGKAAFFEDPHMDN(1297.4379)ESPKYK	1199.1673	17.757
decoy	AMGLLHN(1465.2924)VTRVHELAVKNK	1199.1678	8.9555	decoy	DMGVVFFDRSTAQNAN(1278.4196)DTNSK	1199.1619	11.0005
decoy	NLYRN(1637.6116)RSYFYMQFR	1199.1918	8.7201	decoy	FYEN(1260.2990)KSVRQHMLQLNNWK	1199.1634	8.5615
decoy	DFHFN(1836.6616)KSKTAYNMR	1199.1714	7.5505	decoy	NMFESDLVN(1536.6530)DSRKDCTK	1199.1965	7.3229
decoy	EYYGEFERN(1654.6437)TSQYVR	1199.1785	7.2385	decoy	YTNWQLLKHQMGQN(1261.4319)DSGWK	1199.185	5.9909
decoy	QFN(1694.6892)KTEMMRDTWADK	1199.1906	6.8453	decoy	SHYAHKFNCN(1267.3433)FTMHLRHK	1199.1535	4.5889
decoy	GQTEHVPVQKEPKN(1547.4872)VSDR	1199.1846	6.2557	decoy	N(1296.5354)STTDTNDKDLHSDYCR	1199.1835	4.4139
decoy	QYHANCEMNVN(1677.7382)YSCK	1199.1735	5.7624	decoy	WGYDFVN(1361.4380)YTARKVDLVCK	1199.1863	4.397
decoy	QQFNEWN(1643.5779)MSYAYDKK	1199.1523	4.9396	decoy	NCVKEHVENREN(1388.5277)VTCT	1199.1835	3.5648
decoy	LLMPDVPCSSN(1913.6774)SSFK	1199.1615	4.9396	decoy	ACYMVYN(1575.6397)WSARWVEGK	1199.1916	3.1565
decoy	MPTFSKMN(1887.6664)YTFRGK	1199.1719	4.4829	decoy	WFTHCN(1580.6961)NTDGLDCCSK	1199.1649	2.8399
decoy	N(1755.4980)NSRPRYWFMSLLR	1199.152	4.0451	decoy	VQNCN(1409.4679)WTAQVSQSTDA TTR	1199.1586	2.6244
decoy	NNCQDHRN(1718.6897)YSVSVQR	1199.1837	3.9602	decoy	QN(1391.4678)WTKNWPNEHNFFTLK	1199.1832	2.5942
decoy	HEN(1845.6606)VSNEHTMGPNKR	1199.1613	2.965	decoy	HYLCN(1376.5606)TTEYHTFQEGYR	1199.1788	2.3244
decoy	SFTRN(1604.4920)ATLTGYRSPR	1199.172	2.5651	decoy	VGN(1340.3951)YTGTEAPKVVWVRYR	1199.1829	2.1679
decoy	ESWVKFHQVNN(1474.5269)HSAPGQR	1199.1929	0	decoy	QNFSRN(1578.6032)HSWKNGVVEEK	1199.1938	0
decoy	RVAN(1668.4876)ATGLVYTRYLYTK	1199.1927	0	decoy	KYFVDVGLSN(1518.4152)RTEEYTR	1199.1549	0
decoy	DQAMYAACN(1792.7393)ESMKQR	1199.1719	0	decoy	WFFNQNTYDN(1546.5598)NTTEVR	1199.1614	0
decoy	NCMLN(1831.7240)YTVGWFAK	1199.1696	0	decoy	HKMFYFTNWHN(1470.5188)DTGQAK	1199.1681	0
decoy	HVGFSCKN(1790.6631)GSGQLNGR	1199.1762	0	decoy	SLN(1471.4626)WSAKPDMSTATRMTAR	1199.1699	0

Name	Composition	Mono <i>m/z</i>	Score	Name	Composition	Mono <i>m/z</i>	Score
target	TCILEILN(2303.8409)STK	1199.1827	5.4209	target	LN(1378.4328)GSAPIQVCWYRDGVLLR	1199.2024	2.0622
decoy	ALHFAAGN(2280.7746)VTWK	1199.1615	7.5504	decoy	RSVTNLN(1363.3711)KTGTNPGDEYLP	1199.1742	15.3993
decoy	PWWEN(2291.8156)KSTQK	1199.1577	6.5228	decoy	NWWSCWAN(1182.5144)CTRVEGGASMLK	1199.1975	9.6913
decoy	YCSLN(2314.9140)TSGNHK	1199.1658	5.9909	decoy	APRPCCN(1518.5499)NSRCTQNWR	1199.1556	8.881
decoy	WQQPN(2341.8339)HTKSK	1199.1624	5.126	decoy	QAVVRKLAYYFN(1550.4671)NTAMR	1199.1856	6.6824
decoy	N(2283.9264)TSDQWRYNK	1199.1829	4.2816	decoy	N(1279.4870)DTQEDEPRAPNASVRMNSGK	1199.1883	5.9909
decoy	GDSN(2245.8074)RTPGYKVR	1199.1714	4.1285	decoy	VVLSLCN(1278.3198)CTRDYVYKSPDK	1199.1553	5.2002
decoy	RYGN(2168.8991)QSTNACQK	1199.1876	3.8859	decoy	DCHAAACWWDKVRN(1375.5599)TSDK	1199.1818	4.6675
decoy	YLN(2283.7518)GSTSKWKK	1199.1573	3.5251	decoy	GGGEEQMAVVKYN(1305.5171)ESQEENK	1199.1748	4.6356
decoy	WN(2262.7918)ASADTKQRR	1199.1611	3.1565	decoy	EHNFPGVYCTGN(1356.5390)WSMFHR	1199.1711	4.4829
decoy	N(2145.8925)VTAPSSMDEQDR	1199.1782	2.9745	decoy	GDLCLMYRGVN(1529.6517)TTYDCK	1199.1935	4.1652
decoy	GRWN(2239.6750)LSRVLR	1199.1638	0	decoy	ETDNAHNLEEPSN(1252.4670)LSGNWR	1199.1624	3.5648
decoy	PAEN(2293.9808)CSEPNRK	1199.1952	0	decoy	SCDHAFMPFN(1323.6879)CSGWHCMK	1199.1892	3.5251
decoy	N(2264.8949)WTMNEWHGR	1199.1614	0	decoy	N(1230.5094)MSETQFEFWYMMMYNGR	1199.1595	3.1845
decoy	N(2458.0150)GTQEFQWK	1199.1873	0	decoy	MGN(1535.5851)WTWMKKWQEGYSK	1199.1838	2.3039
decoy	GSHLAKGRN(2252.8488)CSR	1199.1797	0	decoy	AAKAEN(1402.4966)FSYSQKLSCSHHK	1199.1872	2.1122
decoy	YVCN(2246.8510)KTYNMR	1199.1598	0	decoy	PFN(1338.4314)MTGLENKYNDRHQHR	1199.1713	1.7558
decoy	DYN(2392.8995)STFKAEK	1199.161	0	decoy	PN(1379.3610)ASLDPMVVRQEFECVVK	1199.1532	1.7558
decoy	YN(2346.7617)STVVPVKK	1199.157	0	decoy	WCFGN(1356.5501)SSNDWRDLRGNVR	1199.1952	0
decoy	MN(2457.9191)ATVWFLR	1199.1737	0	decoy	VVEWDHGRFDNAGGN(1445.5196)HSPR	1199.1751	0
decoy	GCN(2379.0307)DTVEYMK	1199.1809	0	decoy	EAQEFVMHN(1324.3931)RSQVDLWPGK	1199.1684	0

Table S1. The decoy glycopeptides and their scoring results returned by GPE for the four incorrect glycopeptides that were input as target candidates (labeled in red). The four target glycopeptide compositions are NAAGNFSEPSDSSGAITAR+ [Hex]7[HexNAc]3 (3+, *m/z* 1199.16), WEPPLDDGGSEIINYTLEK+ [Hex]5[HexNAc]3 (3+, *m/z* 1199.19), TCILEILNSTK +[Hex]4[HexNAc]6[Fuc]1[Neu5Ac]1(3+, *m/z* 1199.18), and LNGSAPIQVCWYRDGVLLR+ [Hex]3[HexNAc]4[SO3]1(3+, *m/z* 1199.20) (*N* is the glycosylation site).

CHAPTER V

Absolute Quantitation of Glycosylation Site Occupancy Using Isotopically Labeled Standards and LC-MS

This work has been published by the Journal of the American Society for Mass Spectrometry with reprint permission from the journal.

N-linked glycans are required to maintain appropriate biological functions on proteins. Underglycosylation leads to many diseases in plants and animals; therefore, characterizing the extent of glycosylation on proteins is an important step in understanding, diagnosing, and treating diseases. To determine the glycosylation site occupancy, protein N-glycosidase F (PNGase F) is typically used to detach the glycan from the protein, during which the formerly glycosylated asparagine undergoes deamidation to become an aspartic acid. By comparing the abundance of the resulting peptide containing aspartic acid against the one containing non-glycosylated asparagine, the glycosylation site occupancy can be evaluated. However, this approach can give inaccurate results when spontaneous chemical deamidation of the non-glycosylated asparagine occurs. To overcome this limitation, we developed a new method to measure the glycosylation site occupancy that does not rely on converting glycosylated peptides to their deglycosylated forms. Specifically, the overall protein concentration and the non-glycosylated portion of the protein are quantified simultaneously by using isotope-labeled internal standards coupled with LC-MS analysis, and the extent of site occupancy is accurately determined. The efficacy of the method was demonstrated by quantifying the occupancy of a glycosylation site on bovine fetuin. The developed method is the first work that measures the glycosylation site occupancy without using PNGase F, and it can be done in parallel with glycopeptide analysis because the glycan remains intact throughout the workflow.

5.1 Introduction

N-glycosylation is a common post-translational modification that is closely related to various biological events, including cancer metastasis, viral infection of cells, and antibody-antigen interactions.¹⁻² This modification occurs on an asparagine (N) that is within a consensus sequence of N-X-T/S/C (X could be any amino acid except proline). However, the glycosylation site occupancy depends on the enzymes that catalyze glycan biosynthesis, and the extent of glycosylation can change even for the same glycoprotein that is produced from different cell lines.³⁻⁴ The variability in glycosylation site occupancy is a key indicator of cellular activities, as demonstrated by the correlation between a reduction in glycosylation site occupancy of serum proteins and the severity of congenital disorders of glycosylation (CDGs).⁵ Therefore, it is significant to determine the site occupancy accurately in order to fully understand the impact of protein glycosylation on human health.⁶

In measuring the N-glycosylation site occupancy, the most frequently adopted procedure uses PNGase F to detach the glycan from the protein.⁷⁻⁸ As a result, the glycosylated asparagine (N) is converted to aspartic acid (D) through the PNGase F reaction, inducing an increase in mass of 0.984 Da. This N to D conversion is measured by mass spectrometry (MS), and a larger mass discrimination is achieved by ¹⁸O-labeling of the resulting aspartic acid to facilitate the assignment.^{7, 9} The ratio of the formerly glycosylated asparagine over the non-glycosylated asparagine (and thus the site occupancy) is calculated by comparing the signal of the deglycosylated peptide against the non-glycosylated peptide. This PNGase F method is widely used either to measure the occupancy of partially glycosylated asparagine or to identify novel glycosylation sites.^{8, 10}

Nevertheless, it has been found that chemical deamidation of asparagine could occur spontaneously during sample preparation.¹¹⁻¹² Consequently, in the typical approach that uses the

PNGase F reaction, the non-glycosylated asparagine that undergoes chemical deamidation would be incorrectly assigned as the product of the formerly glycosylated asparagine, which leads to inaccurate quantitation in measuring the site occupancy.¹² Moreover, to quantify the occupancy level by the existing method, the MS signal of the deglycosylated peptide (that contains the aspartic acid) is compared to the signal of the non-glycosylated peptide (that contains the asparagine), and the underlying assumption is that the response factors of these two peptides are the same. However, a recent study indicates the deglycosylated peptide showed reduced signal intensity of up to 50% compared with the non-glycosylated counterpart of equal molar concentration for certain peptide sequences.¹³ Clearly, the currently implemented method of using PNGase F in the quantitative analysis of the N-glycosylation site occupancy has limitations that should be addressed.

Herein we have developed an improved approach for determining the glycosylation occupancy that avoids the disadvantages described above. The key innovation in our strategy is to quantify the natively *non-glycosylated* form of the glycopeptide, using an isotopically labeled internal standard. No glycosidase is added to the sample so that the N-glycan stays intact. Instead, two sets of heavy isotope labeled peptide standards are spiked into the sample before proteolysis, and the digested sample is analyzed by LC-MS. One set of peptide standards is employed to determine the total glycoprotein concentration, while the other standard monitors the non-glycosylated part of the glycoprotein. In this way, the abundance of the glycosylated portion of the protein is calculated by subtracting the non-glycosylated protein abundance from the overall protein concentration, and the site occupancy is then determined. To demonstrate the effectiveness of the PNGase F-free approach we developed, the method was applied to characterize fetuin, which has one partially-occupied N-glycosylation site at Asn-158.

5.2 Experimental Procedures

5.2.1 Materials and Reagents

Four purified synthetic peptides labeled with ^{13}C and ^{15}N on terminal lysine or arginine (denoted as *P1-4, sequences contained in Table S1) were obtained from JPT Peptide Technologies (Berlin, Germany). Bovine fetuin was purchased from Sigma Aldrich (St. Louis, MO) and sequencing grade trypsin was acquired from Promega (Madison, WI). All reagents were of analytical purity or better.

5.2.2 Sample Preparation

A glycoprotein solution of 10 $\mu\text{g}/\mu\text{L}$ was prepared in 100 mM Tris buffer (pH 8.0) containing 6 M urea. The sample was treated with 5 mM tris(2-carboxyethyl)-phosphine (TCEP) and 20 mM iodoacetamide (IAM) in the dark for 1 h at room temperature to reduce and alkylate the disulfide bonds, and 40 mM dithiothreitol (DTT) was added to neutralize excess IAM. Subsequently, the sample was subjected to centrifugal filtration to remove excess urea and DTT using a 10 kDa molecular weight cut-off filter (Millipore, Billerica, MA). The purified sample with a volume of 30 μL was collected and serially diluted by Tris buffer to 0.03, 0.15, 0.6 and 1.5 $\mu\text{g}/\mu\text{L}$. Each solution, containing 75 pmol to 3.75 nmol of protein, was spiked with 50 pmol of the four heavy isotope labeled peptide standards (*P1-4). Trypsin was then added at a 1:30 enzyme-to-glycoprotein ratio, followed by 18 h incubation of the sample at 37 °C. Additional trypsin was added at a 1:100 enzyme/glycoprotein ratio to ensure complete digestion for an additional 4 h at 37 °C. The digestion was stopped by adding 1 μL acetic acid, and samples were stored at -20 °C until analyzed.

5.2.3 N-Deglycosylation

The glycoprotein, 300 μg , was suspended in 30 μL of 100 mM Tris buffer (pH 8.0), and the solution was thermally denatured at 90 $^{\circ}\text{C}$ for 10 min. After the sample was cooled to room temperature, 6 μL PNGase F solution (5000 units/mL, New England Biolabs, MA) was added to the sample, and the mixture was incubated at 37 $^{\circ}\text{C}$ overnight. The deglycosylated sample was subjected to trypsin digestion under the same condition described above except that no isotopically labeled standards were spiked into the sample. The prepared solution was kept at -20 $^{\circ}\text{C}$ prior to the analysis.

5.2.4 LC-MS Analysis

Each sample was analyzed by LC-MS in triplicate. HPLC was conducted on a Waters Acquity UPLC system (Milford, MA), and mass spectrometry was performed on an Orbitrap Velos Pro hybrid ion trap-Orbitrap mass spectrometer (Thermo Scientific, San Jose, CA). Samples (5 μL) were separated using an Aquasil C₁₈ capillary column (320 μm i.d. \times 15 cm, 300 \AA , Thermo Scientific). Mobile phases included eluent A (99.9% H₂O+ 0.1% formic acid) and eluent B (99.9% CH₃CN+ 0.1% formic acid). The following gradient was used: 5% eluent B for 5 min, followed by a linear increase to 40% B in 50 min, and a ramp to 95% B in 10 min. The column was held at 95% B for another 10 min before re-equilibration.¹⁴⁻¹⁵ The mass spectrometer was operated at an ESI spray voltage of 3.0 kV with the capillary temperature of 250 $^{\circ}\text{C}$, and full scan mass spectra (m/z 400-2000) were collected at a resolution of 30,000 at m/z 400. A separate LC-MS experiment was also performed to acquire MS/MS data on two analytes of interest, in which the precursor ions of m/z 1006.20 (eluting at 51.0-51.5 min) and 1006.53 (eluting at 49.5-50.0 min) were selected for collision-induced dissociation (CID) at 35% normalized collision energy, with an isolation width of 3 m/z units.

5.3 Results and Discussion

The workflow for quantitative glycosylation site occupancy analysis is illustrated in Scheme 1. Isotopically labeled internal standards are spiked into the glycoprotein sample prior to trypsin digestion, and the digested mixture is analyzed by LC-MS. For a specific N-glycosylation site, the site occupancy is determined by equation 1:

$$\text{Site Occupancy \%} = \frac{\text{Site-Occupied Protein Concentration}}{\text{Total Protein Concentration}} \times 100 \quad (1)$$

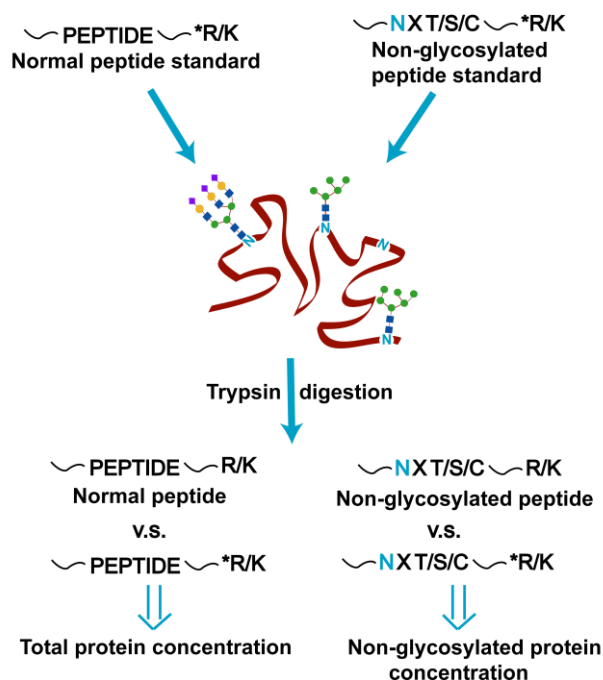
Furthermore, the entire protein population could be divided into two categories: one with the occupied glycosylation site and the other with the unoccupied glycosylation site:

$$\text{Total Protein Concentration} = \text{Site-Occupied Protein Concentration} + \text{Site-Unoccupied Protein Concentration} \quad (2)$$

By combining equation 1 and 2, the site occupancy is calculated by equation 3:

$$\text{Site Occupancy \%} = \left(1 - \frac{\text{Site-Unoccupied Protein Concentration}}{\text{Total Protein Concentration}}\right) \times 100 \quad (3)$$

Accordingly, the total protein concentration is determined by spiking isotopically labeled peptide standards into the protein sample, while the site-unoccupied protein concentration is quantified by using the labeled peptide standard that contains the unoccupied glycosylation site. Therefore, the site occupancy is readily determined without any glycosidase reaction. It should be noted that in order for these equations to be valid, 100% of the partially glycosylated peptide must be accounted for. In other words, if an additional modification were present on the glycosylated peptide, such as a phosphorylation site, this could impact the accuracy of the above-described method. However, these situations rarely occur, and one can verify in advance whether or not other PTMs are present on the peptide containing the glycosylation site to be quantified.



Scheme 1. Peptide standards with heavy isotope labeled tryptic termini (denoted as *R/K) are spiked to quantify the total protein concentration and the non-glycosylated protein concentration.

As a demonstration of the method, the partially occupied glycosylation site of bovine fetuin at Asn-158 was studied.¹⁶⁻¹⁷ As a first step, we verified that no additional PTMs were present on this peptide. Then, three fetuin peptide standards containing heavy isotopes at the C-terminal ends (denoted as *P1-3, sequences listed in Supplementary Table 1) were spiked into the fetuin sample, followed by trypsin digestion. LC-MS data was used to quantify the total concentration of fetuin by comparing the peak areas from extracted ion chromatograms of the fetuin peptides (P1-3) against those of the corresponding standards. A fourth isotopically labeled peptide standard, *P4, was also included in the experiment. This standard was used to quantify the partially non-glycosylated Asn-158 (contained in fetuin peptide P4, VVHAVEVALATFNAESNGSYLQLVEISR, where *N* is the potential glycosylation site), by absolute quantitation of P4 in the same way. Extracted ion chromatograms of the fetuin peptides

and spiked peptide standards are shown in Figure S1; the standards co-eluted with the tryptic peptides of fetuin, as expected.

Table S2 summarizes the glycosylation site occupancy values determined by using the quantitation results of the four fetuin peptides (P1-4), based on different concentrations (0.03-1.5 $\mu\text{g}/\mu\text{L}$) of fetuin spiked with the isotopically labeled internal standards. These data indicate that the glycosylation site occupancy can be measured precisely under different protein concentrations.

The method described in Scheme 1 requires an effective protease digestion because the first three peptide standards are used to determine the concentration of the protein that would be quantified by the fourth standard, if the protein were 100% unglycosylated at the site being studied. In other words, a peptide concentration measured at one part of the protein must be equal to a peptide concentration measured at a different part of the protein. The simplest way to monitor whether or not the peptide concentration is being measured consistently throughout the protein is to use three isotopically labeled peptide standards from different parts of the protein. If each of the three peptide standards produces internally similar results for quantifying the protein concentration, then one can be reasonably assured that the quantitation results of the first three peptide standards are accurately answering the question: How much protein is quantified if the glycosylation site is 100% unoccupied? The fourth standard (P4, which measures the peptide containing the glycosylation site), then, measures the actual (lower) peptide concentration when the glycosylation site is partially occupied.

In order to demonstrate that the first three standards (*P1-3) are effective for determining the concentration expected if 100% of the protein is unglycosylated, each of the distinct peptides (P1-3) are quantified, and their concentrations are compared to the concentration of P4, respectively, to give the individual glycosylation site occupancy values. When the variance of the glycosylation site occupancy is low, as calculated by quantifying different combinations of

peptides (i.e. P1&P4, P2&P4, P3&P4), this implies that the protein is consistently digested at the four isotopically labeled sites and the quantitative result is reliable. As exemplified in Figure 1a, the glycosylation site occupancy values measured by using the three peptide combinations in a fetuin solution of 0.6 $\mu\text{g}/\mu\text{L}$ are internally consistent (ranging from 88.4-90.3%), indicating that the experiment was successful.

To demonstrate what the data would look like when an imprecise, inaccurate result is obtained, we prepared fetuin solutions with reduced trypsin incubation time of 8 h and analyzed the incompletely digested samples under the same workflow. Under these conditions, the four peptides are not expected to be released from the protein in a consistent manner. The resulting glycosylation site occupancy values measured by different peptides, as shown in Figure 1b, vary significantly from each other (ranging from 31.5-90.4%). In this case, one would readily know that the experiment is problematic and the quantitation using any of the three (non-glycosylated) peptide standards is inaccurate. If these results were obtained on an unknown protein, additional attention to the digestion conditions would be needed prior to quantifying the glycosylation site.

In addition to assuring complete digestion, one other experimentally important detail associated with this method is that the MS signals for the peptides need to be measured in the linear response range. To demonstrate that the experiments above were conducted in the linear range, we plotted four calibration curves measuring the instrument response of the peptides across the concentration range used in this experiment. These data are shown in Supplemental Figure 2, and each calibration curve has good linearity (R^2 at or above 0.99).

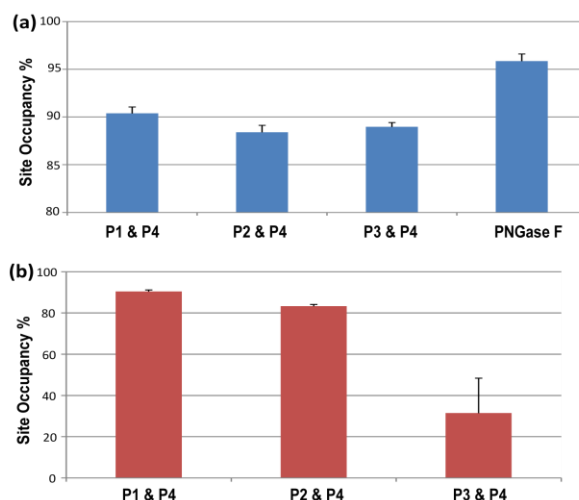


Figure 1. (a) Mean values of percent site occupancy ($n=3$) calculated by using the measured concentration of three different peptides (P1-3) and of the non-glycosylated peptide (P4) in a fetuin solution of $0.6 \mu\text{g}/\mu\text{L}$; the rightmost bar indicates the site occupancy ($n=5$) determined by using PNGase to deglycosylate the protein. (b) The percent site occupancy shown was measured in the same way as (a), but herein the analyzed fetuin samples were incompletely digested.

To compare our new method to the traditional approach, we determined the site occupancy of fetuin using the standard protocol, adding PNGase F to deglycosylate the protein before trypsin digestion then quantifying the percent site occupancy by comparing the peak area of the deglycosylated peptide (m/z 1006.53) to that of the non-glycosylated peptide (m/z 1006.20). The result is included in Figure 1b. In comparison to the approach using labeled internal standards, the PNGase F method results in a higher calculated site occupancy value of 95.8%. We hypothesized that the discrepancy in the measurements is due to inaccuracies in the PNGase F method: Specifically, spontaneous deamidation of the non-glycosylated peptide (P4) is being incorrectly assigned as deglycosylated peptide generated from the PNGase F reaction. If this hypothesis is correct, one would expect to see the spontaneously deamidated peptide even when PNGase F is not present.

The data in Figure 2 demonstrate that spontaneous deamidation is occurring in the sample with no PNGase F added, thus skewing the quantitation results for the PNGase F approach. Figure

2a shows the high resolution MS data of the native non-glycosylated peptide P4 (monoisotopic m/z 1006.2042), and Figure 2b shows the data of the spontaneously deamidated form of this peptide (monoisotopic m/z 1006.5320), which elutes slightly earlier and is heavier in mass by 0.983 Da. Deamidation is also found for the isotopically labeled internal standard (*P4), where the deamidated *P4 (monoisotopic peak at m/z 1009.8675) co-eluted with the deamidated, unlabeled P4, as shown in Figure 2b. The deamidation site can be localized to Asn-158 by comparing the CID-MS/MS data of the peptide P4 (Figure 2c) against the CID data of the deamidated P4 (Figure 2d). As shown in Figure 2c, b-ions (b_6 - b_{14}) and y-ions (y_3 - y_{11}) that do not contain Asn-158 (labeled in blue) have identical m/z values as their counterpart ions in Figure 2d; by contrast, b- and y-ions (b_{24}^{2+} - b_{25}^{2+} and y_{12} - y_{14} , labeled in blue, Figure 2c) that contain Asn-158 are 1 Da less in mass compared to the respective ions (labeled in red, Figure 2d) that carry the Asn-158. Hence we can conclude that chemical deamidation happens on the unoccupied N-glycosylation site, Asn-158. Since the spiked internal standard (*P4) is identical to the non-glycosylated peptide (P4) except for its isotopically labeled C-terminus, it must undergo deamidation to the same extent as the native peptide, assuming the deamidation occurs during sample preparation and not at the protein level. We verified that protein-level deamidation is not occurring by comparing the peak area of the deamidated fetuin peptide (deamidated-P4) to that of the deamidated internal standard (deamidated-*P4). The ratios of these two peak areas were nearly the same as the ratios of the non-deamidated forms of the peptides (P4 and *P4). In summary, the accuracy of our quantitation method is not undermined by chemical deamidation, which induces incorrect quantitative results in the conventional PNGase F approach.

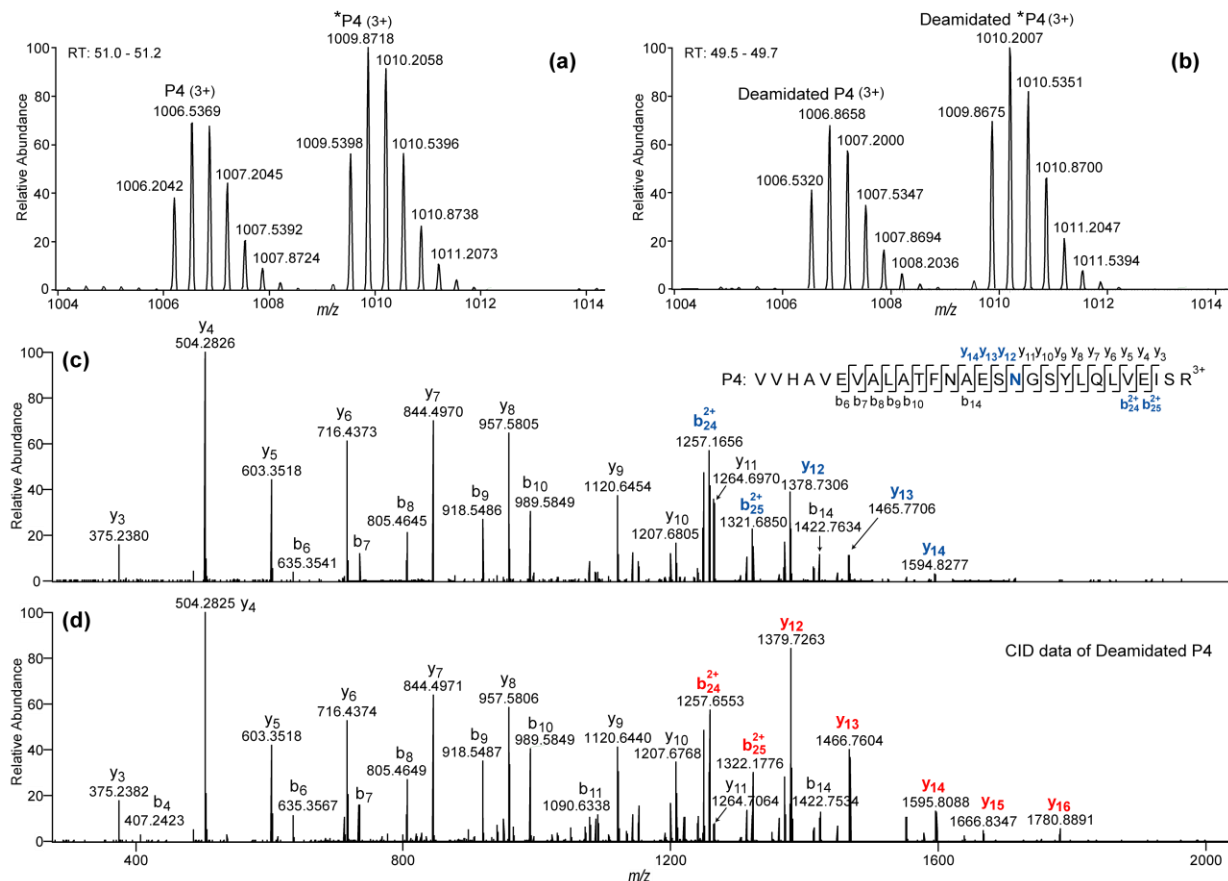


Figure 2. (a) and (b) Zoomed-in mass spectra at retention times of 51.0-51.2 min and 49.5-49.7 min, respectively, of the digested fetuin solution containing heavy isotope labeled standards, with no PNGase F added. (c) and (d) CID-MS/MS data of the non-glycosylated peptide (P4) and its chemical deamidation product (deamidated P4), respectively. In (c), the potential glycosylation site Asn-158 and fragment ions that contain the site are labeled in blue; in (d), fragment ions that contain the glycosylation site are labeled in red.

5.4 Conclusions

We employed stable isotope labeled internal standards in determining the occupancy of a glycosylation site in a protein. The developed method quantifies the overall protein concentration and the amount of the non-glycosylated portion in order to measure the glycosylation site occupancy. No glycosidase is used throughout the protocol. Consequently, the new approach is free from inaccuracies inherent when quantifying using the PNGase F method: Chemical deamidation does not skew the results of the new approach, and one does not need to assume that the deglycosylated peptide and the non-glycosylated peptide have the same response factor. The

presented quantitative method can be easily adopted into typical workflows for glycoprotein quantitation and identification.

References

- (1) Zhou, T.; Xu, L.; Dey, B.; Hessel, A. J.; Van Ryk, D.; Xiang, S.-H.; Yang, X.; Zhang, M.-Y.; Zwick, M. B.; Arthos, J.; Burton, D. R.; Dimitrov, D. S.; Sodroski, J.; Wyatt, R.; Nabel, G. J.; Kwong, P. D. *Nature*. **2007**, *445*, 732-737.
- (2) Drake, P. M.; Cho, W.; Li, B. S.; Prakobphol, A.; Johansen, E.; Anderson, N. L.; Regnier, F. E.; Gibson, B. W.; Fisher, S. J. *Clin. Chem.* **2010**, *56*, 223-236.
- (3) Jones, J.; Krag, S. S.; Betenbaugh, M. J. *Biochim. Biophys. Acta-Gen. Subj.* **2005**, *1726*, 121-137.
- (4) Petrescu, A. J.; Milac, A. L.; Petrescu, S. M.; Dwek, R. A.; Wormald, M. R. *Glycobiology*. **2004**, *14*, 103-114.
- (5) Hulsmeier, A. J.; Paesold-Burda, P.; Hennot, T. *Mol. Cell. Proteomics*. **2007**, *6*, 2132-2138.
- (6) Ivancic, M. M.; Gadgil, H. S.; Halsall, H. B.; Treuheit, M. J. *Anal. Biochem.* **2010**, *400*, 25-32.
- (7) Kuster, B.; Mann, M. *Anal. Chem.* **1999**, *71*, 1431-1440.
- (8) Segu, Z. M.; Hussein, A.; Novotny, M. V.; Mechref, Y. *J. Proteome Res.* **2010**, *9*, 3598-3607.
- (9) Liu, Z.; Cao, L.; He, Y. F.; Qiao, L.; Xu, C. J.; Lu, H. J.; Yang, P. Y. *J. Proteome Res.* **2010**, *9*, 227-236.
- (10) Zielinska, D. F.; Gnad, F.; Wisniewski, J. R.; Mann, M. *Cell*. **2010**, *141*, 897-907.
- (11) Wright, H. T. *Crit. Rev. Biochem. Mol. Biol.* **1991**, *26*, 1-52.
- (12) Palmisano, G.; Melo-Braga, M. N.; Engholm-Keller, K.; Parker, B. L.; Larsen, M. R. *J. Proteome Res.* **2012**, *11*, 1949-1957.
- (13) Stavenhagen, K.; Hinneburg, H.; Thaysen-Andersen, M.; Hartmann, L.; Silva, D. V.; Fuchser, J.; Kaspar, S.; Rapp, E.; Seeberger, P. H.; Kolarich, D. *J. Mass Spectrom.* **2013**, *48*, 627-639.
- (14) Zhu, Z.; Hua, D.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, *85*, 5023-5032.

- (15) Zhu, Z.; Su, X.; Clark, D. F.; Go, E. P.; Desaire, H. *Anal. Chem.* **2013**, 85, 8403-8411.
- (16) Carr, S. A.; Huddleston, M. J.; Bean, M. F. *Protein Sci.* **1993**, 2, 183-196.
- (17) Zhou, H.; Froehlich, J. W.; Briscoe, A. C.; Lee, R. S. *Mol. Cell. Proteomics.* **2013**, 12, 2981-2991.

Table S1. Peptide sequences contained in fetuin and their heavy isotope-labeled peptide standard counterparts. Each standard has ^{13}C and ^{15}N labeled on the terminal lysine or arginine residue. The potential glycosylation site on P4 is labeled in blue.

Peptide	Sequence	Monoisotopic Mass	Charge State	Precursor m/z
P1	EVVDPTK	786.4123	1+	787.4196
*P1	EVVDPT*K	794.4265	1+	795.4338
P2	HTLNQIDSVK	1153.6091	2+	577.8118
*P2	HTLNQIDSV*K	1161.6233	2+	581.8189
P3	EPACDDPDTEQAALAAVDYINK	2405.0693	3+	802.6970
*P3	EPACDDPDTEQAALAAVDYIN*K	2413.0835	3+	805.3684
P4	VVHAVEVALATFNAESNGSYLQLVEISR	3015.5665	3+	1006.1961
*P4	VVHAVEVALATFNAESNGSYLQLVEIS*R	3025.5748	3+	1009.5322

Table S2. The percent glycosylation site occupancy (%) determined by comparing the quantitation results of different normal peptides (P1-3) against the non-glycosylated peptide (P4). Fetuin samples of varied concentrations (0.03-1.5 $\mu\text{g}/\mu\text{L}$) were analyzed.

Concentration ($\mu\text{g}/\mu\text{L}$)	C1 = 0.03	C2 = 0.15	C3 = 0.6	C4 = 1.5
P1 & P4	86.2 \pm 2.8	90.8 \pm 2.5	90.3 \pm 0.7	91.4 \pm 1.0
P2 & P4	86.2 \pm 0.8	88.8 \pm 0.3	88.4 \pm 0.7	91.1 \pm 0.9
P3 & P4	86.4 \pm 2.4	88.4 \pm 0.4	89.0 \pm 0.4	90.2 \pm 0.9

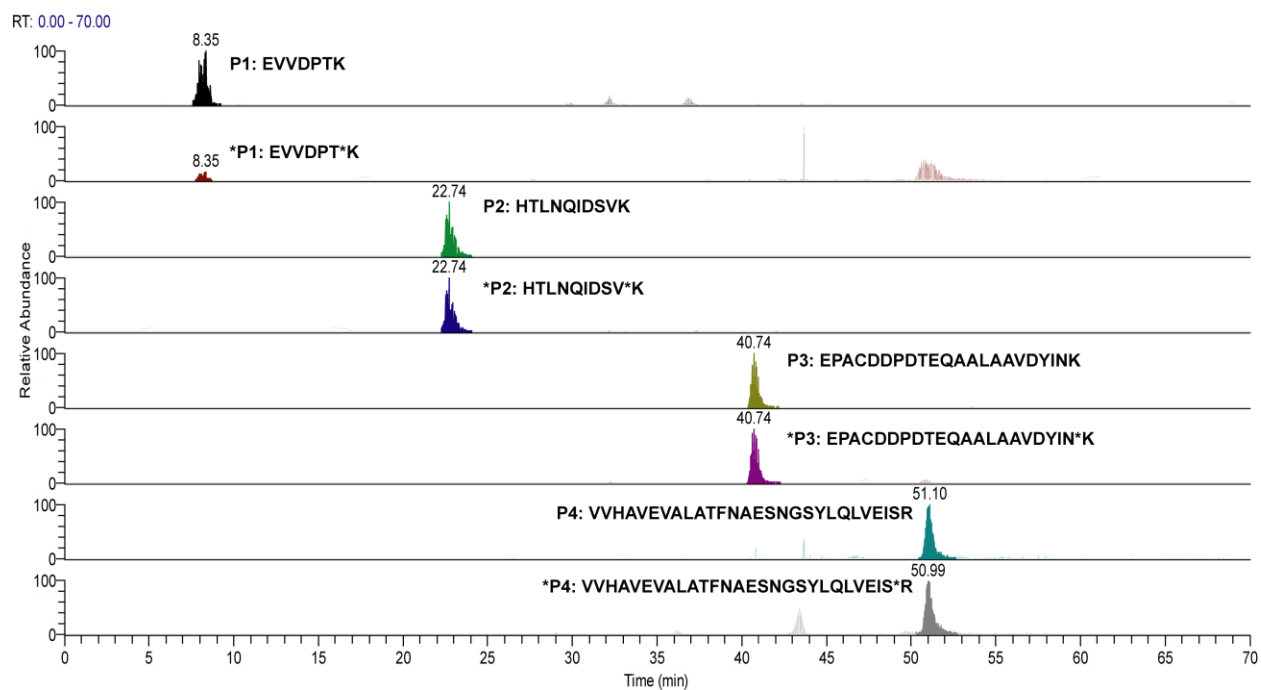


Figure S1. Representative extracted ion chromatograms of the tryptic peptides generated from fetuin and their spiked heavy isotope-labeled standards. The peak area of the monoisotopic peak (m/z values listed in Table S1) of each peptide is used in the quantitative analysis.

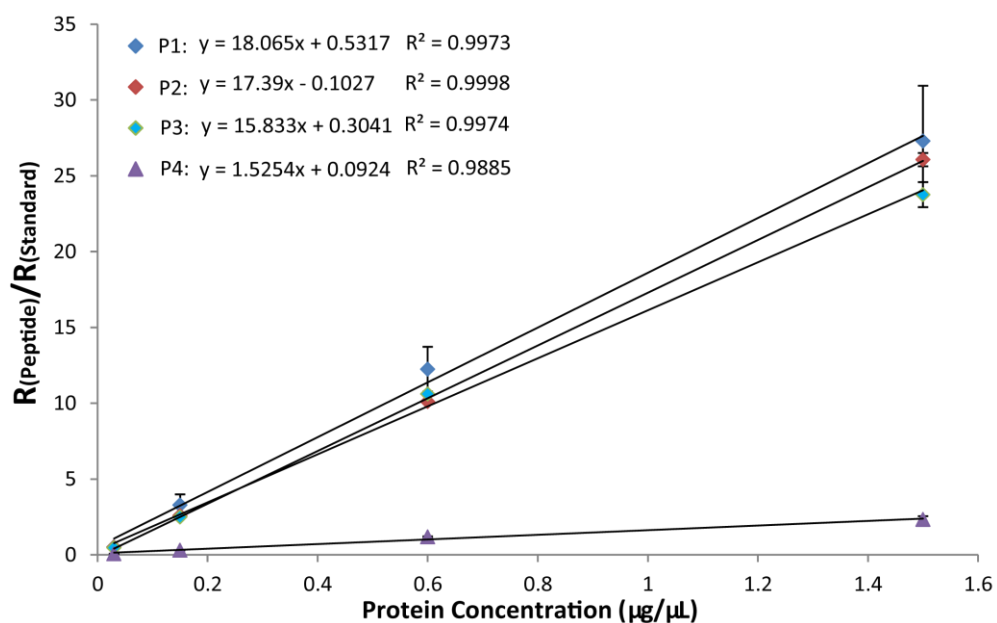


Figure S2. The ratio of each fetuin peptide's (P1-4) signal response over the signal response of the corresponding heavy isotope labeled peptide standard (*P1-4), was plotted against the fetuin concentration (0.03-1.5 $\mu\text{g}/\mu\text{L}$), to construct the four calibration curves shown in this figure. Each sample was spiked with the same amount (50 pmol) of the four heavy isotope labeled peptide standards (*P1-4). Each dilution was analyzed in triplicate.